STAT 213: STATISTICAL MODELING (SPRING 2016)

Instructor. Colin Reimer Dawson (*he/him/his*)
Office. King 204
Email. cdawson@oberlin.edu
Website. http://colinreimerdawson.com/stat213/
Office Hours. General: M/W 4-5:15; Class-specific: T 4:30-5:30 and Th 4-5
Locations and Times. T/Th 11:00-12:15 in King 239

Course Description

Goals. A general goal of the course is to build on the foundational ideas developed in introductory statistics (e.g., STAT 113 or 114), sharpening your ability to reason from data. The theme of the course, as the name suggests, is *statistical models*, which quantify and simplify relationships among variables. Specifically, we will dicuss linear and logistic regression models, and Analysis of Variance models, with arbitrary numbers of quantitative and/or categorical explanatory variables.

Prerequisites / Who This Course is For. This course is designed as a continuation of STAT 113 or 114, and assumes a background in the content of those courses (or an equivalent one) as a prerequisite. Note that the content of introductory statistics at Oberlin recently changed somewhat, and I am sensitive to the fact that not everyone will have taken the most recent version. I expect the main audience of this course to be majors in the natural, social, and computational sciences, who need to be able to create and use statistical models to understand what data can say about interesting questions. Of course, students interested in statistical prediction and inference *per se* should certainly take this course as well. If you are unsure whether this course is right for you, I am more than happy to talk to you about it!

Textbook and Course Outline. The textbook is *STAT2: Building Models for a World of Data*, by Ann Cannon, et al. We will begin with a review of basic concepts that you should be familiar with from your intro course (Chapter 0 in the book). Next we will spend at least half the semester discussing linear regression, where the response/outcome/target variable is quantitative. This corresponds to Theme A in the book, although we will digress to Chapter 5, before returning to Chapters 3 and

Date: Last Revised February 1, 2016.

4, in order to see how the Analysis of Variance (ANOVA) model can be interpreted as a form of linear regression. Next we will discuss the Analysis of Variance in detail, addressing cases with multiple explanatory/predictor variables, and connecting the statistical models to the design of experiments. This corresponds to Theme B. We will spend the last two or three weeks on Logistic Regression (Theme C), which will allow us to model binary response variables with categorical or quantitative explanatory variables. For a more detailed (tentative) schedule, see the course website (link at the top of this syllabus).

Like STAT 113/114, this is a *statistics* course, not a math course, and the focus will be on statistical reasoning, and the use and interpretation of statistical models, not on their mathematical derivations. Some of the mathematical detail will be glossed over, and we will rely on software to do the nitty gritty calculations. Effective use of the computer is an indispensible skill for doing statistics in the 21st century, and constitutes an important part of the course.

Computing. We will use the free and open source statistical computing environment RStudio, which is an interface to the language R. You may either install R and RStudio on your personal machine (`www.r-project.org` and `www.rstudio.com`, respectively), or use Oberlin's RStudio server via a web browser (`rstudio.oberlin.edu`). The R language has become the standard computing tool used by practicing statisticians and data scientists, and so although statistical reasoning is the main goal of the course, competence in R and written presentation of results is a learning objective unto itself as well.

Structure of Class. Most days I will give mini-lectures setting up key ideas, but we will spend a majority of the typical class period working on problems in groups, many of which will involve the computer. There is no dedicated lab day for this class; however, I am hoping that a substantial percentage of you will be able to bring laptops to class on a regular basis, so that there is at least one per group. I will assign groups at random, and we will reshuffle every two to three weeks.

Logistics

Office Hours and Open Door Policy. I will hold some "generic" office hours, and some specific to this class. You are of course welcome to come to either, but may want to opt for the specific hours if possible, since it is more likely that the other students there will have similar questions to yours. You are also welcome to make an appointment, or simply drop by, outside these times (I usually will be in my office during the day Monday through Thursday).

Email. Email is the best way to reach me outside of a face-to-face meeting. You are welcome to address me by my first name, which is generally what I will use when signing emails. I do not consistently respond to email after about 5:30 P.M., but I sometimes check in once later in the evening.

Accommodations. If you have a disability of any sort that may require accommodations in order for you to do your best work in this class, please let me know as early as possible, and consult as well with the Office of Disability Services (ODS). By college policy, *all requests for accommodation require documentation from ODS.*

Honor Code. The Oberlin College Honor Code formalizes the idea that all work that you submit is your own and that you have given credit to the ideas and work of others when you incorporate them. You will be asked to write and sign the honor pledge on each written assignment that you hand in. The honor pledge reads: "I have adhered to the Honor Code in this assignment."

What it means to adhere to the honor code depends on context. For each assignment type, I describe what it means to follow the honor code on that assignment below.

More information about the honor code can be found on the web at the Dean of Students site:
`http://new.oberlin.edu/office/dean-of-students/honor/students.dot`

Graded Work

The grade will be based on two midterms and a final exam (375 pts total), weekly homework and lab problems (225 pts total), three or four data analysis projects (300 pts total), and in-class writing prompts (100 pts total). This adds up to 1000 possible points. I will drop one homework and one lab assignment. Final letter grades will be based on the standard percentages, unless I decide to adjust the thresholds downward somewhat at the end of the term, based on the overall distribution of grades in the class. More detail on each component is given below.

Exams (375 pts). There will be two in-class midterms (**on Tuesday, March 8th and Thursday, April 21st**) and a final exam **on Wednesday, May 11th from 2-4 P.M.** *Honor Code: These must be done individually, but you may use a double-sided hand-written 8 1/2 × 11 note sheet. I don't expect that calculators will be needed, but if it is, I will indicate as much in advance of the exam.*

Homework and Lab Problems (225 pts). I will identify a handful of problems at the end of each class, some of which will be conceptual questions about the material just discussed (the "homework problems") and some of which will involve modeling real data and will hence require the use of the computer (the "lab problems"). There will be one homework set and one lab set due each week, electronically via Blackboard. The homework problems may be prepared using whatever means desired, but should be exported/"printed"/scanned to PDF once finished, to avoid format compatibility issues. The lab problems should be prepared using R Markdown, compiled to HTML, and then exported/"printed" to PDF. For those who have not used R Markdown before, we will go over it. *Honor Code: Each person must prepare their own homework and lab writeups, but you are encouraged to work on the problems together. However, working together may not consist of copying another person's written responses or code.*

Data Analysis Projects (300 pts total). There will be three or four larger scale projects throughout the semester, each focused on a particular major topic of the course (e.g., linear regression, ANOVA, logistic regression). For each one, you have the option of using either a provided dataset and question of interest, or coming up with your own project, either with data you find or data you collect yourself. In either case you will be asked to carry out the four step process of choosing, fitting, assessing, and using a statistical model in order to best answer the question. These projects are somewhat like large lab problems. They should be prepared and formatted in the same way as a lab assignment (see above). *Honor Code: For those doing the provided project, each person is responsible for their own work, and should* not *collaborate with others. For those doing your own project, you have the option of working in a group of 2-3. If you found data, your analysis must be novel.*

In-Class Writing Prompts (100 pts). Each day I will identify a set of readings (mainly from the textbook) that you should do *before* the next class, along with a few basic conceptual questions about the material addressed in the reading. I will ask you to turn in your response to one such question at the start of each class (I will indicate which one in class). If you did not feel like you understood the reading well enough to answer the question, you may instead turn in a question that you have about the reading. At the end of each class I will ask you to write down one thing from that day that you found interesting or confusing, or a question that you have about what we did. These assignments will be graded only for "good faith effort", so for the most part your grade for this component will be the percentage of these that you completed. *Honor Code: These must be done individually in class, though you are encouraged to discuss the reading and the reading questions with each other prior to class.*