STAT 213: STATISTICAL MODELING (SUMMER 2021)

Instructor. Colin Reimer Dawson (*they/them*)
Office. King 204
Email. cdawson@oberlin.edu
Course Website. colindawson.net/stat213/
Class Slack Workspace. stat213s2021.slack.com
RStudio Server. rstudio.oberlin.edu
Locations and Times. MWF 1:30-2:20, King 106 (MW) and on Zoom (F).

Office Hours.
M 2:30pm-4:00pm (Group Office Hour in King 203 – Math Library)
T 11:00am-12:00pm (drop in, King 204)
W 9:30am-10:30am (Zoom, by appointment)
Th 11:00am-12:00pm (Zoom, by appointment)

Important: If your schedule conflicts will *all* of these times, let me know ASAP; I may be able to rearrange some things

Course Description

Overview. 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 allow us to account for the relationships among two or more variables, and make predictions about one variable using one or more others.

Learning Goals. After completing this course, students should

- Understand **how statistical models**, particularly **multivariable models, are used**

Date: Last Revised June 2, 2021.

- Be able to **select, fit, assess, and compare predictive models** for both quantitative and binary response variables, with the aid of standard statistical software
- **Recognize common pitfalls** that occur when building models, and **employ appropriate caveats** when interpreting models
- Be able to interleave the output of statistical software with written explanations and interpretations to **produce informative analyses** useful to a lay reader

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. 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, though our focus will be on application and practice rather than theory. 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*, by Ann Cannon, et al.

The second or third editions (which has the subtitle "Modeling with Regression and ANOVA") are recommended due to improved explanations and additional examples, but you can use the first edition (which has the subtitle "Building Models for a World of Data") if needed. You do not need any electronic supplements.

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 about half the semester discussing **linear regression**, first with a single explanatory variable, AKA "predictor" (which should be familiar), then generalizing to an **arbitrary number of explanatory variables/predictors**. This corresponds to Unit A in the book.

After that we will skip to **logistic regression**, which is Unit C in the book.

Finally, we will return to **Analysis of Variance** (Unit B in the book), and make some connections between these models and linear regression models.

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**. That said, you will need to be(come) familiar with some basic mathematical notation (for example, nested summations, and indexing variables with subscripts), which will be used to describe models and modeling ideas concisely and precisely.

We will do very little hand calculation, instead relying on software to crunch numbers for us. **Effective use of computing tools 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. Most students in this course should have had some experience using R in their intro course; however, some students (e.g., those coming instead from PSYC 200) may not, and even among those who have, experiences (and retention!) will vary widely. Therefore, we will dedicate some class time during the first couple of weeks to a refresher (or introduction!) to R and RStudio.

I will create accounts for each of you (if you do not have one already) on Oberlin's RStudioPro server, which you can access from any web browser at `rstudiopro.oberlin.edu`. You may optionally choose to install R and RStudio on your personal machine as well by visiting `www.r-project.org` and `www.rstudio.com`, and downloading the software there; however, you will still need to use the RStudioPro server to turn in assignments and access solutions, etc.

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.

Friday classes will be held on Zoom and will serve most weeks as a "lab day". However, there may be additional classes when it will be useful to have a laptop available during class, for brief explorations involving the computer.

Structure of Class. This class will involve mixture of a traditional lecture and in-class active learning activities, generally in small groups of 2-3. In some cases these activities will be intended as a chance to practice with material we have previously discussed in lecture; other times we will explore new ideas in groups *before* we discuss them as a class, to get you thinking about them.

Generally I will assign random groups, shuffling every couple of weeks. I recognize that group work, especially with assigned partners, is not most people's favorite thing; however, small group discussion is valuable for learning, and random assignment with periodic shuffling is considered a pedagogical "best practice" for promoting an inclusive and equitable learning environment. That said, **if at any point you feel that the social or intellectual dynamics in your group are problematic, let me know (privately) as soon as possible.**

Logistics

Communication. I have created a Slack workspace for the class, at `stat213s2021.slack.com`, which you will receive an invitation to join (if you haven't already). This workspace serves as a convenient way to organize communication between you, me, and your peers. **Please direct all course-related communication there, rather than email!** I have a much easier time keeping correspondence organized that way and you will get responses much more promptly.

I recommend installing either the Slack desktop client (`https://slack.com/downloads`) or mobile app (available for iPhone or Android), or both, rather than relying on the browser interface, but it's up to you. **I will usually post classwide communications to the `#announcements` channel**, so I recommend **keeping subscribed to notifications to at least that channel**, or at least checking it regularly.

If you have a question or comment that other students might be interested in, I encourage you to **post to one of the classwide channels** rather than PMing me. You might even get a faster response from one of your peers than from me!

I will try to respond to most questions posted on class days by the next class day. **If you need to ask me about something due the following morning, don't wait until the night before!** I have family and parenting responsibilities in the evenings and on weekends; and besides, it's just poor form. If you don't receive a reply from me, don't hesitate to follow up as it may have slipped through the cracks; I won't be offended!

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.

I take the Honor Code very seriously. This means two things: First, I presume a relationship of mutual trust will not try to police your behavior. However, if I do have reason to believe that this trust has been betrayed, I don't take it lightly, and will not hesitate to involve the Honor Code committee in the matter. I will do my best to remind you of the concrete expectations for each assignment as they come up, but it is ultimately your responsibility to make sure you understand them.

More information about the honor code can be found on the web at the Dean of Students site: `https://www.oberlin.edu/dean-of-students/student-conduct/academic-integrity/students`

Assignments and Activities

The course grade is based on the following elements:

- Demonstrating understanding of specific, discrete concepts and content. This is assessed on
  - Weekly problem sets
  - Short weekly take-home quizzes
  - Two midterms and a final exam
  - Two short "data analysis projects"
- Demonstrating an ability to apply and integrate concepts and skills to solve real-world problems using professional quality tools, and to communicate analyses in writing to an untrained reader (20% of grade). This ability is evaluated through the two data analysis projects.
- Good faith effort to complete problem sets in a timely fashion (20% of grade)

Problem Sets. Homeworks will generally be posted on Fridays and due the following Friday by the start of class. Answers should be **turned in electronically** to the designated folder on the RStudio server. I will fetch submissions from there via an automated script and distribute feedback the same way, so it is important that you put your files in the right place!

A subset of the assigned problems on each problem set will be graded by the student grader, and full solutions will be made available.

It is critical that you do the homework problems; the majority of learning takes place by practicing on your own, not by observing others work through problems. I do not count homework for a grade in order to keep stress associated with them minimal and encourage you to attempt to *deeply* process and *understand* what you are working on, rather than aiming for the "correct" answer.

Honor Code. You are encouraged to collaborate freely on homework; however, **what you turn in must be in your own words / your own code**.

Quizzes. Most weeks on Monday I will hand out a takehome quiz, which will be due the following class (Wednesday). Although these are takehome assignments, you are meant to treat them as though you were doing them in class: do them in a contiguous 30 minute window, without reference to notes or textbook, and without talking to anybody else about them.

Quizzes are graded based on demonstrated mastery of a set of "Specific Learning Objectives", or "SLOs". Although the quizzes count toward your grade, their primary purpose is to give you a chance to check the extent to which you understand a particular concept well enough to communicate about it in writing. Individual SLO marks can be replaced by doing "reassessments"; there will always be opportunities to rebound from a less-than-stellar showing on a quiz.

See the "Grading System" document for details.

Honor Code. Quizzes are closed book and closed notes, and must be done individually. You may use a calculator (a cell phone app is fine, but you may not use any other functionality on your phone).

Exams. There will be two major takehome "midterm" exams and one takehome final due during finals week (Saturday 5/16 at 7pm). Like quizzes, exams are graded using Specific Learning Objectives (SLOs). Also like quizzes, SLO marks on the first exam can be repaced by doing "reassessments". There are no in-class exams.

See the "Grading System" handout for details.

Honor Code. You may have one double-sided 8 1/2" × 11" note sheet which is **handwritten by you**, as well as a calculator, for each exam.

Projects. There will be two data analysis projects in the second half of the semester. One will be concerned with linear regression, the other with logistic regression. For each of these, you will be asked to take a dataset and a motivating question or set of questions (I will provide some options, or you can choose something yourself), and build, assess and use a set of predictive models to help you say something about the questions.

See the "Grading System" handout for details on how project grades are computed.

Honor Code. You may discuss your work with anyone, but you must write all code and text yourself (note that taking someone else's text and altering wording does not constitute "writing text yourself"). You must appropriately cite sources of any ideas that you did not originate.