

STAT 213: STATISTICAL MODELING (SPRING 2018)

Instructor. Colin Reimer Dawson (*he/him/his*)

Office. King 204

Email. cdawson@oberlin.edu

Website. <http://colinreimerdawson.com/stat213/>

Office Hours. M 3:30-5:00, W 9:15-10:45, F 3:30-5:20. **Important:** If your schedule conflicts will *all* of these times, let me know ASAP, so I can rearrange some things

Locations and Times. MWF 11-11:50, King 239.

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, including multivariable models, are used
- 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. 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 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 Theme A in the book. After that we will skip to logistic regression, which is Theme C in the book. Finally, we will return to Analysis of Variance (Theme 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, though 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. Effective use of computing tools is an indispensable 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.

There is no dedicated lab day for this class; however, I am hoping that many of you will be able to bring laptops to class on a regular basis, so that there is at least one per group.

Structure of Class. This class is a hybrid between a traditional lecture course and an “inquiry-based learning” (IBL) course. We will spend significant chunk of class time working on explorations in groups, sometimes to practice with material we have previously discussed in lecture, sometimes to explore new ideas to get a feel for them before we discuss them as a class.

You will stay with the same group for a few weeks, and then reshuffle. **If at any point you feel that the social dynamics in your group are problematic, let me know (privately) as soon as possible.**

LOGISTICS

Communication. I am trying out using the Slack app instead of email for course related communication this semester. I have created a group called `stat213s2018` that I will invite you to join. You can then send me, or the whole class, or your group, messages in a dedicated channel. You can download Slack for free on a computer, phone, or tablet, and can be logged in on more than one device at a time. Group discussions will be organized by topic into “channels”, and you can direct message me for private communication.

You should allow about 24 hours for me to respond to most things; it will often be less, but if you need a response by a particular time, be sure to ask me the day before. I am unlikely to respond on Saturdays.

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>

DELIVERABLES AND GRADING

Your Responsibilities. Your responsibilities in this class include the following:

- Attending and being engaged in class (20% of grade). “Being engaged” involves
 - Participating actively in in-class group activities
 - Completing short daily writing prompts about course readings and in-class discussions (10% of grade)
 - Making a good faith effort to do the weekly problem sets (10% of grade)
- Demonstrating understanding of specific concepts and content (60% of grade). This is assessed on
 - Short (10-15 min.) weekly quizzes
 - One in-class midterm exam
 - An in-class final exam
 - Two “data analysis projects”, and one optional final project
- Demonstrating an ability to apply and integrate content 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 and the optional final project.

Writing Prompts. You will be asked to submit two things via Slack by 11:59 P.M. the night before every class:

- (1) (At least) one thing you found particularly interesting or particularly puzzling from the *last* class
- (2) (At least) one thing you found particularly interesting or particularly puzzling from the reading for the *upcoming* class

These can be brief: something like one or two sentences each (though you are welcome to write more if you want). The goals of these prompts are (a) to help me stay attuned to what made sense and what didn't as we go so I can spend extra time on more difficult and more interesting topics, and (b) to give you a concrete bit of motivation to keep up with the reading.

In addition, if at any time you want to give me feedback anonymously on how the course is going for you, or comment on things you think I'm doing that you like or don't like, or things you think I'm not doing that I should, I will provide an anonymous "suggestion box" via a link to a Google form on the course website.

Problem Sets. Assigned homework problems should be typed and "printed" to .pdf (or, in a pinch, handwritten and scanned to .pdf) and turned in via Blackboard by 11:59 P.M. on the posted due date.

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. There will be a quiz most Fridays, excluding the first week, and excluding weeks when there is an exam. Quizzes will be at the beginning of class, and take between 10 and 15 minutes. Quiz grades can be repaced by doing "reassessments". 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).

Extra Time. If you have an accommodation from ODS for extra time on quizzes, it would be best if you can arrive early, since when quizzes end we will go on with other material. If this is not possible due to scheduling constraints, let me know and we will work something out.

Exams. There will be two required in-class exams, one right before spring break (Friday, 3/16) and one on the last day of classes (Friday, 5/11). There will be an *optional* final exam on the designated final exam day for the course. It is optional in the sense that it cannot lower your grade, but it can serve as an extra opportunity to show mastery of some content that you didn't master the first time through. Exam grades can be repaced by doing "reassessments". See the "Grading System" handout for details on grading.

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

Projects. There will be two data analysis projects toward the end of the semester, one on multiple linear regression, with a tentative due date of Monday 4/2, and one on logistic regression, with a tentative due date of Monday 4/23. 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 regression models to help you say something about the questions.

A third, optional, data analysis project can be completed, due the day of the (also optional) final exam. This project can look like either of the first two projects (though should involve new data and questions, chosen by you), or can combine different types of models if using more than one type makes sense for the questions you want to ask. 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.

Final Grade Calculation. The final grade is based on the following components:

- Concepts and content (60%)
- Application, integration and communication (20%)
- Consistent engagement with writing prompts (10%)
- Good faith effort on problem sets (10%)

In-class active participation is expected as a given, but grade-wise only affects subjective "judgment calls" in other contexts (for example, the standard for "good faith effort" is higher if you have not been coming to class, since I will assume this means you feel comfortable enough with the material on your own that you can skip; similarly borderline grades will be influenced by participation).

Each of these 4 components will be scaled to a 0-4 scale and then averaged using the weights above. Exact letter grade correspondences will be decided at the end of the semester based on the class-wide distribution, but letter grades will be no *lower* than the following.

Average (0-4 scale)	Minimum Letter Grade
3.6	A
3.4	A-
3.2	B+
3.0	B
2.8	B-
2.6	C+
2.4	C
2.2	C-
2.0	D

Important Note: Due to the numerous grade replacement opportunities in the course, there will be a general upward trend in running averages. So you should *not* extrapolate from your current grade: it is *assumed* that you will revisit content and your grade will tend to go up as you go along.