STAT 213: STATISTICAL MODELING (SPRING 2020)

REVISED TO REFLECT "VIRTUALIZATION" IN MODULE 2

Times. MWF 2:30-3:20 EDT
Locations. King 239 (Module 1); Zoom (Module 2)
Instructor. Colin Reimer Dawson (*he/him/his*)
Email. cdawson@oberlin.edu
Website. http://colinreimerdawson.com/stat213/

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
- 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 edition (which has the subtitle "Modeling with Regression and ANOVA") is recommended due to its 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 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 Analaysis of Variance (Theme B in the book), and make some connections between these models and linear regression models.

For a more detailed 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. 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, I recognize that some students (e.g., those coming instead from PSYC 200) may not. For those students (or anyone who wants a refresher), I have provided select lab assignments

from my STAT 113 course on the course website which you may want to work through on your own.

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.

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 (Revised).

Module 1 (in person. 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.

Module 2 (virtual classes via Zoom). I will post video lectures at least 24 hours before each class meeting (likely broken into multiple segments) covering the scheduled content for that day. **You are expected to view these and write down questions and comments (include slide numbers and timestamps wherever applicable) prior to the start of each class, and post these to Slack.** We will then spend the class period on discussion, clarifications, additional examples, and in "breakout sessions" where you can work together on homework problems (I will virtually "circulate" around the breakout rooms)

Logistics

Communication. I have created a Slack workspace for the class, at `stat213s2020.slack.com`, which you have received an invitation to join. This workspace serves as a convenient way to organize communication between you, me, and your peers.

I strongly 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.

When contacting me outside class time or office hours, sending me a PM on Slack is the best way to ensure a prompt response. I generally try to respond to Slack messages the same day if they are sent before 8pm, but it may sometimes be the following day. Email may be somewhat slower than that, but is preferable for less time-sensitive correspondence. If you don't receive a reply from me within 24 hours for a Slack message, or 72 hours for an email, don't hesitate to follow up as it may have slipped through the cracks.

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.*

Edit: With the transition to a virtual format for the second half of the semester, I am cognizant of the fact that many of us may be trying to accomplish our work under circumstances that may be significantly less conducive to academic study than our usual on campus community. Please do not hesitate to talk to me if you have needs that may require additional flexibility.

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 extremely 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 treat this as a very serious matter, 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`

Deliverables and Grading (REVISED)

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

- 20%: Being engaged in class. "Being engaged" involves
  - Module 1: Participating actively in in-class group activities
  - Module 1: Completing short daily reflections about course readings and in-class discussions (5% of grade)
  - Module 2: Posting questions and comments on video lectures prior to each class meeting (15% of grade)
- 10%: Completing weekly homework assignments
  - Homework problems are graded based on completion/good faith effort. Some problems will receive feedback on correctness using the SLO grade scale, however these marks are for feedback purposes only and do not factor in to the actual course grade.
- 70%: Demonstrating understanding of specific, discrete concepts and content, and mastery of skills. This is assessed on
  - Short weekly take-home quizzes
  - One "midterm" exam
  - One "data analysis" project

  The content is broken down as follows:
  - Big picture and linear regression concepts (40%)
  - Logistic regression concepts (20%)
  - Practical application and communication (10%)

  Most concepts will be assessed at least twice: once on a quiz, and a second time on either an exam or project.

The linear regression content will be assessed primarily on quizzes before spring break and again on the first exam, with a third opportunity to demonstrate mastery in many cases on the data analysis project or the second exam.

The logistic regression content will be assessed primarily on the quizzes after spring break, and then again on either the second exam or the second project (whichever you choose to do).

The "practical application and communication" content will be assessed on the data analysis project, and then (optionally) again on the second project (if you choose to do it).

Writing Prompts (REVISED). You will be asked to post a brief "reflection" via Slack message to me **by 11:59 P.M. the night before** every class.

These can be brief; about 1-2 sentences each (though you are welcome to write more). 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

(b) to give you a concrete bit of motivation to keep up with the reading

The reflections will take a slightly different form in Module 2, as well as take on increased importance, when we have moved to a virtual classroom format.

Module 1. During the in-person portion of the class, each reflection should include both of the following:

(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

Module 2. During the virtual portion of the class, your reflections should consist of the following:

(1) A brief (1-2 sentence) synopsis of what you took to be the main points of the **video lecture** based on the material for the upcoming class

(2) A brief (1-2 sentence) synopsis of what you took to be the main points of the **reading** assigned for the upcoming class

(3) (At least) one thing you would like to have some discussion about during class time, from either the lecture, the reading or both

Problem Sets. Homeworks will generally be posted on Fridays and due electronically the following Friday by class time. Answers should be **typed** and turned in electronically as instructed for that homework.

Problem sets are graded based on a "good faith effort" to complete all assigned problems to the best of your ability. Feedback will be provided by a student grader on a subset of the assigned problems, 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. The reason I do not count homework for a grade is not that it is unimportant; to the contrary, I see it as particularly important to keep stress associated with them minimal, and encourage you to approach the homework as a *formative* exercise: I want you to attempt to *deeply* process and *understand* what you are working on, rather than simply 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 (/distribute) 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 still 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).

Exam (Revised). There will be one required exam, distributed **Friday, April 3rd** and due back the following **Friday, April 10th**. This exam will cover all of the material on (single and multiple) **linear** regression (but not the material on logistic regression, which we will have started on by then).

Like quizzes, the exam is graded using Specific Learning Objectives (SLOs). Also like quizzes, SLO marks on the first exam can be repaced by doing "reassessments".

See the "Grading System" handout for details.

The second exam is now optional; you need to do *either* the second exam or second project, but not both. If you choose the exam option, it will emphasize logistic regression, but will also include questions on "big picture" ideas that pertain to both linear and logistic regression.

Honor Code (Revised). The exam is open book and open notes, and you can make free use of a calculator and/or statistical software to do computations; however **you may not collaborate with other students**, and **you may not use other internet sources**. Note that, although you are allowed to use R or similar, the exam is not assessing your mastery of software.

Project (Revised). There will be one required data analysis project on linear regression. The project specifications will be distributed on Monday 3/30, and the completed project writeup will be **due Monday 4/20**.

For this project, you will be asked to take a dataset and a motivating question or set of questions (I will provide some options, but I encourage you to choose your own topic that you are personally interested in), and build, assess and use a set of predictive models to help you say something about the question(s).

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

The second project is now optional (you need to do *either* the second exam or second project, but not both). If you choose the project option, it will have the same format as the first project, but will require you to employ logistic regression models. **In either case, the final writeup will be due on the date of the scheduled final, which is Saturday, May 16th, at 9pm**.

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 (Revised). The final grade is based on the following components:

- Concepts, content, and skills (70%)
  - Big picture and linear regression concepts (40%)
  - Logistic Regression concepts (20%)
  - Practical application and communication (10%)
- Good faith effort on problem sets (10%)
- Consistent engagement with writing prompts, Module 1 (5%)
- Consistent engagement with writing prompts, Module 2 (15%)

Each of these 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.

NEW: Pass/No Pass Option. With the college extendeding the option to convert a traditional letter grade to a Pass/No Pass grade until the last day of the semester, I have defined concrete criteria that you may use to lock in a "pass" mark for the course in case you would like to reduce your workload.

Students who complete Exam 1, Project 1, Homework through HW6, Quizzes through Quiz 5, and keep up with Slack posts through at least

the end of the classes on logistic regression (roughly Wednesday 4/22) **will be guaranteed a grade of "pass" if they have at least a "B" average at that point** (a 3.0 in the SLO grading scale).