STAT 340: Applied Regression Methods

About the Course

Instructor Marie Ozanne
Email: `mozanne@mtholyoke.edu`
Office: Clapp 403
Office Hours: I will hold regularly scheduled office hours each week at times to be selected by you. These times will be posted on the Moodle site. Please do not hesitate to contact me to set up appointments for additional office hours at any time!

Classes The class meetings times are Monday 9:45-11:00 AM, Tuesday and Thursday 9:30-11:15 AM, Wednesday and Friday 10:15-11:15 AM. We will not always use the entirity of class for formal class time; some time will be set aside for questions. Classes will be held on Zoom - the links can be accessed on Moodle.

Course Website Course materials and class forums can be accessed on Moodle. The site will be regularly updated with lecture notes and materials used in class. Lab assignments we will work through in class and homework assignments will be distributed on GitHub; links to these materials will be accessible through Moodle.

Description In this course, we will build on the ideas you have developed in introductory and intermediate statistics courses to develop a set of statistical methods and computational tools that can be used for common data analysis tasks in the frequentist framework. We will take the "applied" part of the course title seriously, focusing on working with real data sets to answer real questions.

In terms of statistical modeling, we will plan to cover the following material:

- Review of simple and multiple linear regression; matrix formulation of regression; polynomial regression
- Generalized linear models (GLMs), including general formulation, logistic, poisson, and negative binomial regression
- Generalized linear mixed effects models (GLMMs)
- Time series
- Spatial modeling

In terms of statistical computation, we will use R and cover the following material:

- Data management in R's tidyverse
- Data visualization with R's ggplot2
- Version control and collaboration with git and GitHub

Textbook The primary text for this class will be "Applied Regression Analysis & Generalized Linear Models" by John Fox. This is available as a hard cover (ISBN 978-1-4522-0566-3). I may occassionally assign readings from "An Introduction to Statistical Learning with Applications in R" by James, Witten, Hastie, and Tibshirani (available for free at `http://www-bcf.usc.edu/~gareth/ISL/`) and from "R for Data Science" by Wickham and Grolemund (available in website format at `http://r4ds.had.co.nz/`), or other open source resources.

Time commitment As you all know, we will be completing this class in 7.5 weeks. Typically in a regular semester, you are expected to budget approximately three out-of-class hours for every credit hour to complete readings, assignments, and homework (12 hours per week of time commitment). In the module structure, you should budget approximately six out-of-class hours for every credit hour to watch recorded lectures and complete other assignments (24 hours per week of time commitment). If you are spending more time than this on a regular basis, please check in with me.

Policies

Synchronous/Asynchronous Classes Classes will be delivered synchronously via Zoom 3-4 days per week (more typically 3), and usually will be a combination of discussion and group work. On the other 1-2 days, you will meet asynchronously at a pre-specified time with your assigned small group (organized by time zone) to work through labs/other activities. This is intended to give students in disparate time zones some flexibility, while still satisfying requirements for synchronous instruction. Asynchronous days will be Wednesday, and sometimes Friday.

Attendance I realize this term is going to come with challenges, likely both physical/mental health and technology related. That being said, the material will move quickly due to the compressed nature of the module and it is important that you attend virtual classes as frequently as possible. If anything happens that prevents you from attending class or working with your group for more than a day or two, please let me know as soon as possible so that we can come up with a plan to ensure your success in the class. You can go into whatever level of detail makes you feel comfortable.

Collaboration Much of this course will operate on a collaborative basis, and you are expected and encouraged to work together with a partner or in small groups to study, prepare for quizzes, complete homework assignments, and prepare presentations. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or large blocks of R code from another student is not acceptable and will receive no credit or a penalty. No interaction with anyone but the instructor is allowed on any quizzes. All students, staff and faculty are bound by the Mount Holyoke College Honor Code.

To sum up: On homeworks and labs, **I want you to work together**. *But,* **you must write up your answers yourself.**

Cases of dishonesty, plagiarism, etc., will be reported.

Technology

Computing with R Modern statistics can't be done without computation. We will use the R statistical programming language in this course. R is one of the most commonly used programming languages in academic statistics, and I use it daily; it's also very commonly used in statistics and data science positions in industry. Knowing R is a marketable skill. In this class, you will use R nearly every day, and for many homework problems. I expect that you are familiar with R from previous classes, but I do not expect that you are an expert at R yet. That said, it is imperative that you let me know if you are confused about anything we are doing in R.

We will use R via RStudio; Mount Holyoke's version of RStudio Server can be accessed at `https://rstudio.mtholyoke.edu/`. If you are off campus, you must access the server through Mount Holyoke's VPN. You are also welcome to work locally on your own computer if you have RStudio set up; however, please make sure you have installed at least version 3.5.0 of R and the latest versions of any R libraries we use.

Version Control with Git and GitHub Git is a version control system that facilitates working on coding and writing projects collaboratively, and allows you to revert your code to a previous version if you realize that you made a mistake. Version control systems such as git are used in most modern data science and statistics positions in industry. Part of my goal as an educator in the statistics program is to ensure that you are prepared to enter the work force, and for that reason the basic use of git is a learning objective for this course. This means that all labs and the computational portion of homework assignments will be distributed to you in git repositories and submitted by committing and pushing the completed assignment to GitHub. I will provide further details and walk through this process, as well as basic interaction with git, in class. Note that we will use the graphical interface to git that is built into RStudio rather than the command line interface to git.

Assignments

Your grade for this course will be a weighted average of scores from several components:

Item	Weight
Participation and Labs	15%
Homework	40%
Quizzes	5%
Oral Presentation 1	20%
Oral Presentation 2	20%

Participation and Labs The best way to learn statistics is to do it. This class will be built around a series of labs that we will do in class. Although I will not grade these labs for correctness, I expect you to complete them and submit your work on Moodle. I will occasionally look at submitted labs to see how everyone is doing and whether there are any points I need to address in class. I am always happy to answer any questions you have about these labs. A large component of participation will come from participating in discussions and posting questions, answers, and comments in our Moodle forums (details will be given weekly).

Homework We will have regular homework assignments to be completed outside of class. Homework assignments will be short written papers (approximately 5 pages, not including figures or references) detailing various statistical analyses for assigned real world problems/data sets. As we work through the material and homework assignments in this class, our writing and communication of statistical analyses and results will improve. To reflect this process, homework assignments will be graded in a "scaffolded" fashion. This means that, while you will get feedback on your entire assignment, both in the form of peer review and instructor feedback, you will only be formally assessed on portions of the paper, as detailed in Table 1 for each assignment. By the time you turn in your final paper for the semester, you will be assessed on the entire paper, but you will have had plenty of practice to improve your writing along the way. See the homework section of the rubrics posted on Moodle for an assessment rubric for these assignments. **Homework will be submitted to Gradescope, and will be due on Friday of each week.**

Assignment	Section(s) to be assessed
Paper 1	Methods
Paper 2	Methods, Results, References
Paper 3	Methods, Results, Discussion, References
Paper 4	Introduction, Methods, Results, Discussion, References

Table 1: Homework assignment assessment guide

The grading on writing assignments in this class is structured such that you can learn how to write an applied statistics paper and improve your writing skills, without being heavily penalized before you have feedback on the process (Table 1). Even if you do not start out as well as you would like, this setup should allow everyone in the class to improve as statistical communicators. Due to time constraints, I will not be able to assess resubmitted work, but I am happy to answer any questions about the feedback you receive. I expect improvement over the course of the semester, which will be reflected in grades - someone who does these assignments well in the beginning and someone who is still developing, but becomes a strong writer over the course of the module could conceivably receive the same course grade.

Late Assignments While I will post due dates for assignments, there is flexibility in these due dates, so in theory you can turn any assignment in late, without penalty. Because of the pace of a module, however, routinely turning in assignments more than three days late, will make it very difficult to keep up in the class. Homework is due on Fridays - you can take the weekend to either catch up, or get ahead, depending on how your week went.

Quizzes There will be occassional quizzes on material from posted lectures. These are low stakes assignments intended to measure comprehension. Quizzes will be administered through Moodle.

Group Oral Presentations There will be two group oral presentations, one in week five, and one in the last week. Students will be put into small groups and will teach us all something new, either by extending the statistical methods used in one of the homework assignments (e.g., implement a more appropriate method; compare two approaches; account for missing data), or choosing another method of interest from the textbook. Each group will put together an $n \times 10$

minute oral presentation (where n=number of group members) with slides and present it to the class during Week 5 of the module; each group member will be expected to present for approximately 10 minutes. Through these presentations, students will (1) investigate a new statistical method or evaluate the merit of several methods we have learned in class and highlight the strengths and weakness of these approaches; (2) hone both oral presentation and collaboration skills. While the default format for this type of presentation is likely a live virtual presentations, other formats may be acceptable if you have time and want to be creative, like recording a video presentation/instructional video. I will provide a list of possible topics, but other ideas are welcome, as long as they fall into the general framework of regression. Groups will clear their presentation choices with me no more than one week before the presentation.

Accommodations

Academic Accommodations AccessAbility Services is the office on campus that determines academic accommodations for students with disabilities. If you need official accommodations through AccessAbility Services, you have a right to have these met and kept confidential. Please contact AccessAbility Services, located in Mary Lyon Hall 3rd Floor, at 413-538-2634 or accessability-services@mtholyoke.edu. If you are eligible for academic accommodations, you will be provided with an accommodation letter. Once you receive your accommodation letter, I would like to meet with you and discuss these approved accommodations and our class. For more information on who might be eligible for accommodations and the application process please see the AccessAbility Services website (`www.mtholyoke.edu/accessability`).

Religious Accommodations In support of our religiously diverse student population, students may miss a class, obtain an extension on an assignment, or reschedule an exam if there is a conflict with a religious high holiday or observance. Students should **notify me at the beginning of the module if a religious observance will require special accommodation**.

Audio/Visual Recording Policy To encourage active engagement and academic inquiry in the classroom, as well as to safeguard the privacy of students and faculty, no form of audio or visual recording in the classroom is permitted without explicit permission from the professor/instructor or without a letter from AccessAbility Services, signed by the faculty member, authorizing the recording as an accommodation. Authorized recordings may only be used by the student who has obtained permission and may not be shared or distributed for any reason. Violation of this policy is an infraction of the Mount Holyoke Honor Code and academic regulations and will result in disciplinary action.