SDS/MTH 291: Multiple Regression

Spring 2018

Professor Benjamin Capistrant, bcapistrant@smith.edu

Student Hours: Mon 10-12, Wed 2-3, Fri 9:30-11:30
Office: Burton Hall 215

Class Sessions: Tue/Thu 10:30-11:50AM
Class Room: Sabin-Reed 301

Course Description

In this course, students will learn how to use statistical models to analyze real-world data. Students will develop an understanding to how to choose the best regression model for a different types of response variables, include and transform different types of explanatory variables, and fit models with statistical software. This course will encourage and foster communication skills to convey results precisely and clearly to a general, non-technical audience.

The main goal of this course is to learn how to build statistical models - models that will let you understand and estimate the relationships between multiple phenomena, assess differences in an outcome of intereste between multiple groups, and even make predictions. The world we live in is often complicated, full of possible competing explanations for some apparent relationship. Multiple regression moves beyond many of the statistical tests and approaches from introductory statistics that consider simple differences between two variables (i.e., one response and one predictor variable, respecitvely). Instead, multiple regression offers one way to adjust statistically for myriad alternative explanations of the apparent relationship between the response and predictor; or to account statistically if the relationship between the response and predictor varies by other factors. With this course, you will develop useful quantitative modeling skills that allow you to harness data to enrich your understanding of the world we live in.

Course Objectives

After completing this course, you will be able to choose different statstical modeling approaches (and understand their assumptions), be able to fit these models empirically using statistical software (`R`), assess how well the model you fit, and use that model to address your hypothesis, including to communicate what you found to a wide audience. Specifically, the course objectives are for you to:

- Master the 4-step process of choosing, fitting, assessing, and using statistical models
- Identify and fit models for quantitative and categorical responses, including effective use of transformations when conditions are not met
- Understand confounding conceptually, and successfully implement approaches to adjust for confounding and moderation/modification in statistical software
- Detect outliers and influential points in models
- Test hypotheses and construct confidence intervals for slopes of regression models
- Construct prediction intervals based on regression models
- Compare different regression models and identify the best model empirically
- Understand how multicolinearity and correlated predictors affect a model
- Use odds and odds ratios to understand categorical response variables

Prerequisites

A college-level introductory statistics course (e.g. MTH/SDS/PSY 201, GOV 190/203, SOC 203, MTH 219, MTH/SDS 220, ECO 220), a score of 4 or 5 on the AP Statistics Exam, or permission of the instructor. Students need to be familiar with basic statistical concepts including graphical and numerical descriptive statistics, the normal and t-distributions, hypothesis tests, and confidence intervals.

Course Readings

This course uses one textbook. This book is **required** and is available at the Grecourt bookstore in the basement of the Campus Center for $166. In case you choose to acquire the book elsewhere, note that you will not need online access for the purposes of this course.

- Cannon AR, Cobb GW, Hartlaub BA, et al. STAT2: Building Models for a World of Data. New York: W.H. Freeman; 2013.

Additional recommended readings, such as applications of material covered in class in the news, will be available on the course website and via Slack.

Assignments

The course assignments will include a homework assignment and quizes most weeks of the course, two exams, and a group project with sections/drafts due throughout the course culminating in a final presentation and paper. The homework and the project are the backbone of this course and must be completed satisfactorily to receive an A or B grade in the course.

Homework

Problems (mainly from the book) will be assigned (almost) every week on Thursday in class and on Moodle; homework will be due via Moodle before 11:55PM of the following Monday. Homework should be completed in `RMarkdown`; both the `RMarkdown` and resulting pdf files should be submitted. Homework will be graded, and you will be occasionally be asked to present your solution to a homework problem to the class. ***Late homeworks will have 25% off if submitted within 48 hours of the deadline and no points thereafter***. *Everyone has one free 48 hour extension that you may take whenever you need it. Please just let me know when you plan to take it.*.

Final Project

A significant, culminating activity for this class will be to conduct a statistical investigation on a question of interest to you. These projects will be done in groups of three of your choosing. You will use existing data - publicly and freely available data online, or from faculty research you may be involved in. The project will be broken into multiple component parts due throughout the semester, culminating in a 10 minute oral presenation (with slides) in the last week of class and a final written report due the last day of class. We will spend time in class discussing questions of a reasonable scope for this project, reviewing available data sources, and writing a proposal and resulting report.

Tests

There will be three in-class tests throughout the semester. More details about the format and content of the exams will be available closer to the exam. You may use one double-sided sheet of notes and a

calculator for reference during the exam; otherwise, they are "closed book" and no additional materials may be used during the exam. You may not use your phone as the calculator. Project presentations during the last week of classes will take the place of a final exam.

Quizes

There will be brief ($\leq$20 minutes) quizes on Moodle due every week on Monday before 11:55PM (to be precise). The goal of the quizzes is to give timely feedback on your understanding of the previous week's material before the course moves on to new material. In rare cases, a quiz will also include questions about the assigned readings for the week ahead. Quizzes will be easy if you have been keeping up with the reading and the homework. The quiz contents will be discussed in Thursday's class. ***There will be no make-up quizzes****, but your lowest quiz score will be dropped.* You will have one attempt and 20 minutes to complete the quiz once you begin. Quizes will open Friday at noon (12:00pm) and close Monday before 11:55PM; you have 3 1/2 days to find 20 minutes to complete the quiz.

Grading

Course grades will be a weighted average of the grades on your project, tests, homework, and quizzes. A small amount of your grade for participation includes class attendance, participation in class activities and discussions, as well as professional conduct. Professional conduct expectations include: treating classmates, guest speakers, and instructors with respect; ensuring that all communication is done in a respectful and courteous manner - this includes email.

Project & Presentation	25%
Homework	25%
First Exam	20%
Second Exam	20%
Moodle Quizzes	5%
Participation	5%

The purpose of grades is to provide formative feedback that aids your learning. I keep course grades in the online gradebook on Moodle so you can always check them there. But what matters is learning. I will enthusiastically talk to you about your learning anytime, and I encourage discussions in which we go over the work you have completed. These conversations let me hear about your challenges and questions, and provide important learning opportunities. However, my rule is that we shouldn't talk (and especially haggle) about the points or letter grade assigned unless I have made a clerical error.

Resources

Course Website

Course information, including this syllabus, will be available on the course Moodle site. I will update information posted there regularly with class handouts, homework, project assignments, and announcements.

Computing

We will use the `R` statistical software package extensively and exclusively. `R` is open source software that is available for free on Mac, Windows, and Linux operating systems. There is an extensive section of the

course Moodle site with R resources, which may be especially useful for those newer to R. There are two ways to access R for this class:

- **Server:** The Smith College RStudio server, which you are strongly encouraged to use. Enrolled students will get an account for the server shortly after the beginning of the class. The server is located at (http://rstudio.smith.edu). The advantages of using the server version are that it is cloud based: your work will be automatically saved and backed up, and you can work from any computer that has an internet connection (note: you need to be on a Smith College IP address to access the server, to access it off-campus, you may need to establish a VPN connection to Smith). Like any shared resource, there are limits on how much data you can store (only really relevant for projects, not homeworks) and how quickly it processes during periods of high use.
- **Desktop:** You can also download R and RStudio to your computer and run things locally. Note - homework will be submitted via PDF, which requries additional software (TEX). For Mac users, I recommend MacTeX, and for Windows users, I recommend MiKTeX; see more at https://www.latex-project.org/get/.

Note that you do not have to choose one or the other; indeed, you can use both at different points in the class. For instance, you may want to get started using the server for homeworks and then transition to working on your project locally to maximize processing power and speed. Keeping your files on a cloud (e.g., Google Drive, Dropbox) in a folder shared among your group members will facilitate collaboration if you're all working from your individual desktops.

Student Hours

I am available in my office Mondays 10:00AM-12:00pm, Wednesdays 1:00-2:00PM, and Fridays 9:30-11:30AM for students to stop by and discuss anything that relates to the course or would be helpful to talk about in person. These hours are a great opportunity to clarify issues that have come up in class and in homeworks (including R), to discuss research opportunities, your academic interests as they relate to SDS, talk about a news story you read that involved statistical modeling, even just to say hi. Some questions are easier to answer in person than via email, and many - myself included! - feel more comfortable asking questions one-on-one than in front of class. Student hours are a great option for both scenarios. Practically, the hours on Monday are a good opportunity to resolve questions that came up in the previous week's class before the quiz or on the homeworks that are due most Mondays.

Additional Support on Campus

Stat TAs

There are Stat TAs available from 7 to 9 pm on Sunday - Thursday evenings in Burton 301. Each TA has a brief bio at the Spinelli Center website. Most are familiar with the material covered in this course and all can provide support for working in R.

Data Assistants & Data Research and Statistics Counselor

Spinelli Center for Quantitative Learning (now in Seelye 207) supports students doing quantitative work across the curriculum. The Center has a Statistics Counselor and Data Assistants available for appointments.

Writing Support

The Jacobson Center for Writing (Seelye 307) offers a variety of services design to help students improve writing skills, including drafts of their papers. This resource may be especially useful for the final project paper.

Course Policies

Attendance

Your attendance in class - including being on time - is crucial to your and your classmates' success. We are all going to learn this material together, so we need to have everyone present and working. I will make accommodations for an unavoidable absence if you notify me in advance. Our Honor Code means that you will be the judge of whether or not an absence was unavoidable. (For instance, staying in bed because you had the flu would be an unavoidable absence, but oversleeping because you stayed up late to write a paper would be an avoidable absence.) One necessary absence during the semester is not unusual; having more than two is uncommon.

Collaboration

Much of this course will operate on a collaborative basis, and you are expected encouraged and to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or blocks of `R` code from another student is not acceptable and will receive no credit. All students are bound by the Smith College Honor Code.

Written and Oral Communication Skills

Your ability to communicate results, which may be technical in nature, to your audience, which is likely to be non-technical, is critical to your success as a data analyst. The homework and project assignments in this class will place an emphasis on the clarity of your writing. In addition, there were will be many opportunities throughout the class where you will be asked to present your results from in class exercises, homeworks and the final project to your classmates. Consider this course as a way to develop these crucial communication skills and be quick to seek support from peers, in office hours, and from other resources (see above) as needed to do so.

Email and Course Communications

Student Hours and Slack will be the best way to reach me. I am actively on slack during the work day (9-5pm) and sometimes in the evenings. If you have questions about HW, please post them to the Slack channel for that assignment and answer other peoples' questions if you know the answer! If you do get in touch via email, I will do my best to respond within 24 hours during the week and 48 hours on the weekend. It will be helpful if you include an informative subject line, including the course number (i.e., "SDS 291 HW 3 Question").

Timeliness, Deadlines, and Extensions

Deadlines matter in this class, both because these parameters help us keep our focus on the work rather than logistics and because they matter in "the real world". When assignments are due *before* 11:55pm, Moodle counts submissions at 11:55:01pm as late. This policy can be very challenging for people who put

off their work until the last minute; I encourage you to start homework assignments early. Late work has 25% off if submitted within the first 48 hours after the deadline (for homework, typically before Wednesday by 11:55pm). I allow for one "free", no-penalty 48 hour extension. If you need another extension or a longer extension during the semester, I grant extensions only if your Class Dean is involved well in advance of the assignment deadline and suggests that flexibility with deadlines would be useful.

Academic Honor Code

The Smith College honor code articulates general expectations for students' conduct throughout the course. While some specific applications of the honor code are noted throughout other sections of the syllabus, these examples are illustrative and not exhaustive. In accordance with the honor code statement below, suspected cases of academic dishonesty will be reported to the Academic Honor Board.

> Students and faculty at Smith are part of an academic community defined by its commitment to scholarship, which depends on scrupulous and attentive acknowledgement of all sources of information and honest and respectful use of college resources.
>
> Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations. All submitted work of any kind must be the original work of the student who must cite all the sources used in its preparation.

Again, every word that you write must be your own. Copying and pasting sentences, paragraphs, or blocks of `R` code from another student is not acceptable, will receive no credit, and warrants notification of the Academic Honor Board. Similarly, Moodle and self-scheduled examinations should be completed independently and should not be discussed with other students during the open exam/quiz window. Per College policy, suspected violations will be reported to Academic Honor Board.

With this in mind, I note that many of the materials used in this course build substantially from previous course instructors' work, namely Ben Baumer, Amelia McNamara, Miles Ott, Chad Topaz, and Jude Higdon. I thank them for their work and willigness to share these materials.

Course Accomodations

If you have documented accomodations with Office of Disability Services, please share the letter from their office via email as soon as possible. This includes if changes to exisiting accomodations or any new accomodations emerge through the semester.

Health and Well-Being

Your well-being matters to me personally and as a course instructor vis-a-vis your success in this course. I encourage you to use your support resources (including those on campus, like your Class Dean, RA, and Health Services) early and often. If you have social, health, emotional or financial challenges that significantly affect your performance in the course, please feel free to email me and set up a time to talk.

Course Schedule

The following tentative outline lists readings, homeworks, quizes and other deadlines for each week of the class. Please complete the reading assignments *before* coming to class so that you can participate fully in the discussion. I reserve the right to revise this schedule – updates will be posted on Moodle.

Topics Schedule

Week	Dates	Topic
0	1/22-26	Intro, review terms and notation. Simple Linear Regression models, conditions
1	1/29-2/2	SLR: Estimation, slope inference, ANOVA table, confidence intervals
2	2/5-9	SLR: SLR review / Confounding, Intro to Multiple Linear Regression
3	2/12-16	MLR: comparing two regression lines, coding categorical predictiors, second-order models
4	2/19-23	MLR: multicollinearity; F-tests and Nested F tests
5	2/26-3/2	MLR: Bootstrap for regression, Leverage & Influence
6	3/5-9	MLR: Review and Summary, First Exam
7	3/12-16	Spring Break
8	3/19-23	Topics: Data Wrangling, Visualizations
9	3/26-30	MLR: In Class Practice, Additional Topics in MLR
10	4/2-6	LOG: model and fitting
11	4/9-13	LOG: odds ratios & assessing model, multiple predictors, LRT test
12	4/16-20	Topics: Multiple Comparisons, Missing Data, Transformations
13	4/23-27	LOG: Third Test, Project Work Time
14	4/30-5/4	Project Presentations
15	5/7-5/11	Finals Week - Project Reports Due

SLR: Simple Linear Regression. MLR: Multiple Linear Regression. LOG: Logistic Regression.

Assignments & Deadlines

Week	Dates	Mon	Tue	Wed	Thu	Fri
0	1/22-26				0	
1	1/29-2/2	HW1, Q1	0, 1		2.1-2.2	
2	2/5-9	HW2, Q2	2.3-2.5		3.1, 3.2	*Roster*
3	2/12-16	HW3, Q3	3.3		3.4	*Draft Proposal*
4	2/19-23	HW4, Q4	3.5-3.6		3.7-3.8	
5	2/26-3/2	HW5, Q5	4.3		4.5	*Final Proposal*
6	3/5-9	HW6, Q6	0-3		**First Exam**	
7	3/12-16	[Spring Break]				
8	3/19-23		Reading TBD		Reading TBD	*Data Appendix*
9	3/26-30	HW7, Quiz 7	Reading TBD		Reading TBD	
10	4/2-6	HW8, Q8	9.1-9.2		9.3-9.5	*Draft Results*
11	4/9-13	HW9, Q9	10.1-10.3		10.3-10.5	
12	4/16-20	HW10, Q10	7.2		Reading TBD	*Paper Draft*
13	4/23-27	HW11, Q11	Review/Project Work Time		**Second Exam**	
14	4/30-5/4		*Presentations*		*Presentations*	
15	5/7-5/11					*Final Paper*

HW: Homework, Q: Quiz. *Project assignments in italics.* Campus holidays in [brackets]. **Tests in bold**.