STA 371H: Statistics and Modeling (Honors)

Course Syllabus

Spring 2018

Course overview

Course Details

Section 1: 04160
Time: M W 9:30 AM –11:00 AM
Place: CBA 4.330

Section 2: 04165
Time: M W 11:00 AM –12:30 PM
Place: CBA 4.330

Instructor: James Scott
(james.scott@mccombs.utexas.edu)
Office hours: M W 2:00-3:00 PM
Office: CBA 6.478

Teaching Assistant: David Puelz
(david.puelz@phd.mccombs.utexas.edu)
Office hours: TBA
Location: TBA

`jgscott.github.io/STA371H_Spring2018`

This course will teach you how to address hard questions in the face of uncertainty and randomness, using the tools of data science and probability. What policies contribute the most to creating sustained economic growth? Do charter schools work miracles for their pupils, or benefit from self-selection? Do "green"-certified commercial properties command a premium on the rental market, all else being equal? What balance of stocks and bonds should go into your retirement portfolio when you're 22? (How about 52?) We usually cannot study questions like these using controlled experiments. Rather, we must carefully sift through a body of evidence, hoping to tease out relationships in complex, noisy data sets that don't look anything like something we might draw up in a laboratory.

In this course, you will learn to analyze data and use it to make better decisions. My goal is to help you cultivate three important skills:

1. Using visual and quantitative evidence to evaluate hypotheses in loosely structured problems with no verifiably correct answer.
2. Building simplified models of real-world systems to aid cause-and-effect reasoning and guide intelligent behavior.
3. Using knowledge of probability to improve your decisions.

You'll encounter lots of real data sets—from finance, politics, culture, sports, marketing, economics, science, and so on. By semester's end, you will have learned some lessons that will serve you well throughout the rest of your life, both as a data analyst and a data consumer. The main focus of this course is on data, not math. This does not mean that we won't encounter some math along the way—but as you'll soon discover, the math behind data science is surprisingly simple. Even if you've never thought of yourself as a math person, please don't worry. I promise you'll be up to it—and I also promise it will be worth it.

Materials

Readings

You do not need to purchase a textbook: I have always hated the idea of assigning an expensive shelf decoration that is, at best, a halfway match with the goals of the course. Instead, we will have three main references, all free:

1. An online course packet, available from the course website: `jgscott.github.io/STA371H_Spring2018`
2. The first five chapters of *Statistical Modeling: A Fresh Approach*, by Daniel Kaplan. These five chapters are available for free; I have provided links on the course website.
3. *Data Analysis for Politics and Policy*, by Edward Tufte. Out of copyright, with a PDF copy freely available on the course website.

In addition to these resources, there will also be shorter readings, all of which will be linked from the course website when the time comes.

Software

For software, we will use an open-source statistical programming environment called R (www.r-project.org) for almost of what we need to do. R is freely available for Mac, Windows, and Linux. It's the real deal—a software package used by organizations as diverse at Google, J.P. Morgan, Whole Foods, Facebook, and the New York Times to analyze their data. I want you to come away from the course with a legitimate, industrial-strength platform for data analysis. Specifically, we will use a graphical front-end to R called RStudio: www.rstudio.org. (You will almost surely like this better than the standard Mac or Windows interface to R.)

Detailed instructions for installing and getting started with R are available from the course website.

Course outline

The following day-by-day outline is subject to review if we need to slow down or speed up. But it should give you a more detailed idea of how the course will proceed. There are 25 topics that must fit in 28 class days, not counting holidays and the mid-term exam day. This gives us some leeway for extra days to devote to examples, or to simply slow down. You'll also see that the topics are split into two overall

sections: statistical modeling (roughly the first half of the course) and probability/decision-making (roughly the second half). Another important major idea of the course—Monte Carlo simulation—comes up in both halves.

Important dates:
March 8: In-class midterm exam
March 13–17: Spring Break (no class)
April 7: Projects due
May 3: Last class meeting

Statistical modeling

1. Data exploration.
Topics: Principles for plotting data. Numerical and categorical variables. Contingency tables. Simple summaries and graphics: histogram, boxplot/dotplot, scatter plot. Numerical measures of dispersion. Variation between and within groups. Multivariate plots.

2. Fitting equations by least squares.
Topics: Fitting straight lines via ordinary least squares. Interpreting the model parameters. Basic stories one can tell with a statistical model: prediction, summary, adjustment, quantifying information.

3. Beyond straight lines.
Topics: Transforming and combining variables. Fitting nonlinear models by adding polynomial terms. Exponential growth and decay.

4. Power laws.
Topics: Fitting power laws using the double log transformation.

5. Predictable and unpredictable variation.
Topics: Prediction intervals. The decomposition of variance; R^2.

6. Grouping variables as predictors.
Topics: Dummy variables and interactions in group-wise models. The analysis of variance.

7. Numerical and grouping variables together.
Topics: Dummy variables and interactions in models involving a numerical predictor together with grouping variables. Dependence among predictors.

8. Parameter uncertainty and the sampling distribution.
Topics: The fundamental frequentist thought experiment. Sampling distributions and standard errors. Bootstrapped approximations to sampling distributions, standard errors, confidence intervals. The Gaussian regression model.

9. Introduction to multiple regression.

Topics: Linear models with more than one numerical predictor. Interpreting partial slopes. Statistical adjustment.

10. Multiple regression, continued
Topics: Isolating partial relationships. Bootstrapping multiple regression models. Grouping variables in multiple regression. Assessing statistical significance. Prediction intervals from multiple regression models.

11. Testing hypotheses: the basics.
Topics: Simple binomial tests. Permutation tests. Test statistics. *p*-values.

12. Testing hypotheses in regression.
Topics: Hypothesis testing in multiple regression models. Assessing statistical significance for grouping variables.

13. Building a predictive model.
Topics: Out-of-sample predictive validation; cross-validation. AIC and stepwise selection.

14. Cause and effect.
Topics: Experiments; natural experiments; matching; regression.

Probability and decision-making

15. Introduction to probability.
Topics: The NP rule; the NNT in medicine; basic rules of probability.

16. Conditional probability.
Topics: Probability distributions. Joint, conditional, and marginal probabilities. Rule of total probability and aggregation paradoxes.

17. The compounding rule.
Topics: Winning streaks; compounded probabilities; the fallacy of mistaken compounding

18. Bayes' rule.
Topics: Updating conditional probabilities

19. Probability models 1.
Topics: Simple discrete probability distributions. Expected value and variance. Law of large numbers.

20. Probability models 2.

Topics: The Binomial and Poisson distributions.

21. Probability models 3.
Topics: The normal distribution. Using Monte Carlo simulation for complex probability distributions.

22. Correlated random variables.
Topics: Regression to the mean; the bivariate normal distribution.

23. Simulation case study.
Topics: Simulating correlated outcomes. Estimating value at risk.

24. Regression for discrete outcomes.
Topics: Link functions. Logistic regression.

25. Data collected in time.
Topics: Trends; seasonality; using lagged predictors.

How the course is structured

On a day to day basis, this course revolves around independent inquiry, in addition to traditional lectures. The focus of class time is on building your capacity to think about open-ended problems. On any given day, there will be a mix of lecture, discussion, and hands-on modeling.

The upshot of all this? You will end up learning many important skills outside of class. This learning will take (at least) three forms:

Reading. Mostly this will be out of the online course packet, but there will also be supplemental readings from other sources.

Videos and software modules. These take the form of both short lecture videos and web pages that walk you through the basics of performing analyses in R.

Practice. There will be weekly problem sets, consisting mostly of open-ended modeling problems. You are encouraged, but not required, to work on these in groups of four people or fewer. We will spend a lot of time working on these problems in class.

You'll find links to all the relevant material through the class website.

As you might imagine from this description, succeeding in this course will require substantial time devoted to out-of-class preparation. As a rule of thumb, you should expect to spend 3 hours per week in class; 1–3 hours per week reading and completing the software modules; and anywhere from 3-6 hours per week completing the exercises.

As with all college classes, in some weeks there will be less than this, and in some weeks there will be more.

Prerequisites

The formal university prerequisites for this course are: Business Administration 324 or 324H; Management Information Systems 301 or 310; Mathematics 408D, 408L, or 408M; and Statistics 309 or 309H. Note: the calculus prerequisite is not there for show. In particular, I expect you to remember derivatives and basic material on logarithms.

Exams and grading

Grades will be determined by a midterm (20%), final exam (30%), project (25%), and homework/in-class quizzes (25% total).

Grading
Midterm: 20%
Final exam: 30%
Homework/quizzes: 25%
Project: 25%

Homework and quizzes

Homework is assigned weekly, and will count for 25% of your final grade. All homework must be turned in *as a hard copy* at the beginning of class on the day it is due (no electronic submission). You are allowed (but not required) to work on homework in groups of 4 people or fewer. If you work in a group, turn in a single write-up, with all your names on it. Homework is graded on a 10-point scale. Homework will be accepted only in hard-copy form. No electronic copies will be accepted without prior permission received before the homework is due. I will grant such permission only in extenuating circumstances.

Homework: assigned weekly and due in hard-copy form at the beginning of class on the due date.

On occasion, we will have unannounced in-class quizzes. These will involve short, simple "did you the do the reading?" kinds of questions. Each quiz will count the same as a single homework assignment.

No late homework will be accepted, and no makeup quizzes will be offered, for any reason. At the end of the semester, I will drop two homework/quiz grades, effectively allowing you two missed homework assignments or quizzes without penalty.

Project

There is one course project, which you may complete in a group of four people or less if you wish. (As with homework, groups are optional.) This will involve getting your own data set on a question that interests you, running an appropriate statistical analysis, and writing up your conclusions. The details will be discussed in class, and posted on the

course website, but the idea is for you to work on the project during the three weeks immediately after Spring Break (a time of the class when you will have no additional homework assignments). This project is worth 25% of your grade, and is due by 5 PM on Friday, April 6, 2018.

Project 1: due Friday, April 6, 2018.

Exams

The in-class midterm will take place on the last day of class before Spring Break. You will need a blue book and an ink pen, and nothing else (no notes, calculator, laptop, etc.).

Mid-term: in class on the last class day before spring break: Wednesday, March 7, 2018

The final exam is worth 30% of your grade and will take place during the usual University exam period in early May. As with the midterm, you will need a blue book, an ink pen, and nothing else. I will announce the time slot as soon as it is available from the University registrar.

Final
When: TBA, during official exam period
Where: TBA

Missed exams

You will not be allowed a make-up for a missed exam without a documented and verifiable medical excuse, or documentation that a family emergency prevented you from attending. The only documentation I will accept for this purpose is an electronic or written letter from Student Emergency Services in the Office of the Dean of Students notifying me of your absence. The Dean of Students will, in turn, require supporting documentation from you (e.g. a doctor's note or letter from primary care provider) in order to verify your illness, injury, or emergency. While this policy may seem strict, it is the only way we can be fair to everyone.

If you will be out of town representing the University on an academic, athletic, or student-organization trip, you must speak with me and provide me with appropriate documentation at least 2 weeks in advance. I will be glad to make arrangements for you to take the exam before you leave for your event.

Finally, if you must miss an exam for the observance of a religious holy day, inform me at least 2 weeks before the exam, so that alternative arrangements can be made in conjunction with the Dean and the relevant university offices.

If you miss an exam for any other reason—including personal travel or family holiday—you will get a zero.

Re-grade requests

On occasion you may notice a simple clerical error in the recording of a grade, which I am happy to correct without hassle. Other regrading requests must be submitted in writing within 7 days of the marked paper being returned. Keep in mind that the entire paper will then be subject to re-grading, and that your grade may go up or down as a result.

Attendance

There is no explicit attendance component to your course grade. Having said that, it is hard to do well on the rest of the course assignments without coming to class. You will also get a zero for any in-classes quizzes you miss—no exceptions. I encourage you to form good attendance habits.

Curving grades

The raw percentage scores to the right will guarantee you *at least* the corresponding grade.

I reserve the right to curve grades up. But I will never curve them down. That means these grades are a floor, not a ceiling, on the final grade that someone with the corresponding raw score would receive. The precise details of any curve are at my sole discretion, and if I should choose to use a curve, I will detail the cutoffs used when course grades are submitted.

Percentage	Grade
93–100	A
90–92	A-
87–89	B+
83–86	B
80–82	B-
70–79	C
60–69	D

Other course details

Quantitative reasoning flag

This course carries the Quantitative Reasoning flag. Quantitative Reasoning courses are designed to equip you with skills that are necessary for understanding the types of quantitative arguments you will regularly encounter in your adult and professional life. For more details, see `www.utexas.edu/ugs/core/flags/quantitative-reasoning`.

Classroom etiquette

You are expected to participate in class; close and put away your laptops, unless it's a "hands-on" day where I ask you to look at data and run models in class; and to turn off your electronic devices. I also ask

that you arrive on time to class, since late arrivals disrupt things for all other students. In turn, I will make sure we finish on time so that students may reach their next lectures/hot dates.

Cheating, plagiarism, and such

Acts of academic dishonesty are ethically wrong; they harm the reputation of the school and demean the honest efforts of the majority of students. Additionally, you should consider three things:

(1) Cheaters are a tiny minority. The vast majority of students who preceded you did it the honest way. Follow their lead.
(2) You play like you practice. The habits you form now will predict the headlines that people write about you, or your company, later in life. Try Googling "Jeff Skilling" or "Fabulous Fab" if you don't believe me.
(3) If you cheat, you're playing with fire. The minimum penalty will be a zero for that assignment or exam. You also risk failing the course and being dismissed from the University.

The bottom line when it comes to cheating is: just don't do it. You might fool me, if you're very lucky and very unscrupulous. But you are highly unlikely to fool the McKinsey interviewer you were hoping to impress with your knowledge of statistics. And you may find that the job market is far more ruthless than university judicial boards.

Now for the usual boilerplate. The McCombs School of Business has no tolerance for acts of scholastic dishonesty. The responsibilities of both students and faculty with regard to scholastic dishonesty are described in detail in the BBA Program's Statement on Scholastic Dishonesty.[1] By teaching this course, I have agreed to observe all faculty responsibilities described in that document. By enrolling in this class, you have agreed to observe all student responsibilities described in that document. If the application of the Statement on Scholastic Dishonesty to this class or its assignments is unclear in any way, it is your responsibility to ask me for clarification. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since dishonesty harms the individual, all students, the integrity of the University, and the value of our academic brand, policies on scholastic dishonesty will be strictly enforced. You should refer to the Student Judicial Services website[2] to access the official University policies and procedures on scholastic dishonesty as well as further elaboration on what constitutes scholastic dishonesty.

[1] http://www.mccombs.utexas.edu/BBA/Code-of-Ethics.aspx

[2] http://deanofstudents.utexas.edu/sjs/

Student privacy

First of all, you should know that I am legally barred from discussing your course performance with anyone other than you and anyone that you explicitly designate. That includes your parents.

Second, a note on Canvas. Canvas is a password-protected web site, and is created automatically for all accredited courses taught at The University. Site activities could include exchanging e-mail, engaging in class discussions and chats, and exchanging files. In addition, Canvas includes a class e-mail roster. Students who do not want their names included in such an electronic class rosters must restrict their directory information in the Office of the Registrar, Main Building, Room 1. For information on restricting directory information, see `www.utexas.edu/student/registrar/catalogs/gi02-03/app/appc09.html`.

Students with disabilities

Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259 `http://diversity.utexas.edu/disability/`.

Campus safety

Please note the following recommendations regarding emergency evacuation from the Office of Campus Safety and Security, 512-471-5767, http://www.utexas.edu/safety.

- Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside.
- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation should inform the instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors.
- Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- Behavior Concerns Advice Line (BCAL): 512-232-5050

- Further information regarding emergency evacuation routes and emergency procedures can be found at: http://www.utexas.edu/emergency.