Math 401: Statistical Modeling

Spring, 2016

Course Description: Data on almost any topic is constantly at your fingertips. Much of this data is high dimensional. We will develop tools to create, test, and use mathematical models based on such data. We study numerous models, emphasizing the conditions for for each, how to build good models, and how to test models. A weekly one-hour lab component demonstrates course topics via real-world applications, using the statistical computing language R to carry out the necessary computations. There will be a mid-semester project and a semester-long project where students can apply their knowledge to a data set of their choosing.

Course Goals:

(1) Practice how to break a complex problem into simple pieces. Develop your analytic thinking skills and problem-solving skills. Learn to mold theory-based approaches to solve messy real-world problems.

(2) Learn to manipulate data sets in both R and Excel. Come to appreciate the power of simulation. Learn skills of data visualization, expectation, estimation, inference, and prediction.

(3) Develop group work skills, test-taking skills, and writing skills. Become proficient at re-writing and communicating technical content to a non-technical audience.

(4) Develop time management skills, metacognitive skills, and the habit of thinking intentionally about your learning and your goals.

(5) Learn about many well-studied statistical models, how to tell which to apply in a given situation, and how to develop new models.

Instructor:	David White
Office:	Olin 202
Extension:	6644
Email:	david.white@denison.edu
Class Meetings:	9:30-10:20 MWF, 9-9:50 Th, in Olin 217
Office Hours:	10:20-11:20 M, 11:30-12:20 W, 10-11 Th, 1:30-2:20 F, **and by appointment**, Olin 202
Final Exam:	Monday, May 9, 9-11am in Olin 217
Web Resources:	http://personal.denison.edu/~whiteda/math401spring2016.html

Your Final Grade. Your final grade in this course will be calculated according to the following:

Homework and labs:	35%	Quizzes:	15%
Midsemester and Final Projects:	15%	Midterm Exam:	10%
Participation:	5%	Final Exam:	20%

Course Overview This course will be broken roughly into five pieces:

- Review of STAT1 via randomization based inference, bootstrapping, and non-parametric statistics.
- Linear Regression, detecting failures of conditions, Logistic Regression.
- Multivariate Regression, Poisson Regression, Negative Binomial Regression, and Analysis of Variance (ANOVA, MANOVA, ANCOVA)
- Dimension Reduction Techniques (principal component analysis, factor analysis), Fitting Distributions.
- What to do when hypotheses fail to hold (Time Series Analysis, weighted least squares regression).

Text and Materials. The textbook that you will need for this course is Stat2: Building Models for a World of Data by Cannon, Cobb, Hartlaub, Legler, Lock, Moore, Rossman, and Witmer, published by W.H. Freeman.

Keys to Success

- Prior to exams, be able to solve every quiz and homework problem under time pressure. Have a perfect, hand-written copy of each of the homework exercises assigned from the book. Use this to study for exams.
- Review the material from class the same day it is given. Find a way to attach this new knowledge to things you already understand. Study statistics a bit every day rather than in bursts just before an exam.
- Focus note-taking in class on these connections you made to your unique prior knowledge, not on material already in the book, handouts, or slides.
- Read the textbook slowly and carefully, at your desk, with a notebook nearby to write down questions. Have R open while reading. Type code in, play around, and experiment to figure out what R can do. Pay particular attention to examples, and try to work them out yourself before reading the solution. Type in and run all code given in the book!
- Start labs and homework early. Give yourself time to get stumped and to get past these difficulties. Focus first and foremost on making sure you understand the data set and can compute on it.
- Keep a list of key definitions, statistical tests, assumptions, and R functions you frequently use, and commit them to memory throughout the course. Test your memory each week.

Homework will be due approximately once per week, but will be assigned every day. Homework is a mix of problems from the textbook, small writing exercises, and exercises in R and Excel. Many of the exercises in R will be completed using DataCamp, and you will submit screenshots to prove you did the requested tutorial. In addition, you are expected to keep up with the reading. In my opinion, the

best way to read a math textbook is to skim before class and try to understand which concepts might be difficult for you *and why*. This will help you get the most out of class. After class, do an in-depth reading of the chapter and complete the homework problems, then skim for the next day. You should expect to spend about 10-12 hours on coursework outside of class per week. Exercise numbers and pages to read will be posted daily on the course webpage.

Collaboration on homework is strongly encouraged, but you should write up your homework yourself, and it should be well written and readable. You should not email code to each other, or share code electronically. However, you are allowed to discuss code with each other and help each other debug. I am happy to answer questions in my office hours, and I encourage you to come if you are confused about anything. You will get the most out of this time if you attempt the homework first and come with questions already prepared.

Labs feature real-world data sets, train students in how to model using statistics, and frequently require written lab reports demonstrating understanding of the model, why it fits, and how to use it. On Thursdays, we'll learn the value of simulation, computational methods for coping with large amounts of data, and how to clean data. We'll use applets to help visualize key concepts, and we'll use DataCamp as an interactive way to develop skills in R. Labs will always be due the following Friday, i.e. 8 days after our lab meeting. No late labs will be accepted. In consideration for sickness, personal emergencies, etc. I will drop the lowest lab grade.

Mid-semester project: Early in the semester you and a partner of your choosing will obtain data on a topic you'll choose in consultation with me. This will be a dataset on which you plan to run a multiple linear regression or logistic regression analysis. This is similar to the final project from Math 242: you will write a paper discussing where the data came from, conducting your analysis, and reporting your conclusions. In addition, you will present a poster and we'll have a day for poster presentations. Your paper will be written in RMarkdown.

Semester-Long Project: Early in the semester you and a partner of your choosing will obtain data on a topic you'll choose in consultation with me. This will be a data of more complexity than the mid-semester data set. Each time we learn something new you'll interpret it or test it on your pet data set, so it should exhibit sufficient complexity to do multivariate regression, logistic regression, MANOVA, and either PCA, time series analysis, or weighted regression. You'll come up with questions about this topic you hope to answer, and may need to obtain more data as the course progresses. At the end of the semester you'll hand in a longer paper answering your questions and give a presentation of your findings to the class.

Quizzes and Exams: There will be a comprehensive 2 hour final exam May 9, 9-11am in Olin 217. To help you consolidate what you have learned we'll have several quizzes during the semester. Longer quizzes (more heavily weighted) will be announced on the course webpage, but many classes will start with a surprise

short open-note quiz on the reading. There will be a midterm exam on Thursday, March 10, 7:00-9:00pm. **Please mark your calendar!**

Course Format. The course meets 4 days per week, for 50 minutes each day. Class will begin with a Q & A forum where I will attempt to clear up any confusion you may have about the reading. Please take advantage of this and come with questions prepared; in class we will often not review the reading unless you ask questions, opting instead to work out examples of the types of considerations you learned about from the reading. Thursday will often be a lab day, so you can sharpen your skills in R.

Communication. It cannot be stressed enough how essential communication is to succeeding in this course. After identifying topics that may be giving you trouble, please communicate this information to me. There's no such thing as a bad or unwelcome question. Additionally, please communicate with each other. Explaining concepts and examples to each other is a great way to learn. It is my goal to create a comfortable environment best conducive for learning.

Participation. Since working with other students is a major part of this course, it is important that everyone participate. Class attendance is therefore essential. Each day you'll be graded on a scale from 0-3, with 0 signifying an absence, 2 attending attentively, and 3 active participation such as asking or answering a question.

Grading Scale: A standard 10% grading scale will be used: 60% is required to pass the class, 70% will be a C-, 80% will be a B-, and 90% will be an A-.

Disability: Any student who feels he/she may need an accommodation based on the impact of a disability should contact me privately as soon as possible to discuss his/her specific needs. I rely on the Academic Support & Enrichment Center in Doane 104 to verify the need for reasonable accommodations based on documentation on file in that office.

Academic Integrity: Academic dishonesty is, in most cases, intellectual theft. It includes, but is not limited to, providing or receiving assistance in a manner not authorized by the instructor in the creation of work to be submitted for evaluation. This standard applies to all work ranging from daily homework assignments to major exams. Students must clearly cite any sources consulted, including classmates who have been collaborators on the homework and online sources of aid. Neither ignorance nor carelessness is an acceptable defense in cases of plagiarism.

I expect that you will all abide by the honor code in this course. Please do not use resources outside of me, your fellow students, the tutors, and the textbook. For the R portion on homework, labs, and projects, you may search for R code online, but you must cite the source for any code you use. Please do not search for solutions to written homework, quiz, or exam questions online. Exams will have a laptop portion where you can showcase your knowledge in R, but you may not use anything else on your laptop, including the internet, txt files, pdfs, etc. If you are

caught with any files open other than RStudio, you will be reported to the Board of Academic Integrity.

Collaboration on homework and projects is permitted, but you should acknowledge in your write-up when you gave or received help on the assignment. Collaboration on quizzes and exams is not permitted. Violations of the honor code will be reported, and violations may result in failure in the course, suspension, or expulsion.

Appropriate Use of Course Materials: The materials distributed in this class, including the syllabus, exams, quizzes, handouts, study aides, and in-class presentations, may be protected by copyright and are provided solely for the educational use of students enrolled in this course. You are not permitted to re-distribute them for purposes unapproved by the instructor; in particular you are not permitted to post course materials or your notes from lectures and discussions online. Unauthorized uses of course materials may be considered academic misconduct.

Email: I will frequently contact you via email. Please check your email regularly. I will also check my email regularly, but often not after 8pm.

Topics: A rough schedule of topics follows, but is subject to change. The schedule on the course webpage supersedes the schedule here:

Weeks 1 and 2: Review of Statistics 1, randomization based inference, nonparametric statistics.

Week 3: Special topics from regression: transformations, outliers, Cook's Distance

Week 4: Logistic regression

Week 5-6: Multivariate, polynomial, Poisson, and negative binomial regression

Week 7: Multivariate logistic regression

Week 8: Quiz, Exam, Mid-Semester Projects

Week 9: Break

Weeks 10-11: ANOVA, MANOVA, ANCOVA

Week 12-15: Time Series Analysis, Principal Component Analysis, Factor Analysis, Weighted Regression

Week 15: Big Data, Final Presentations.

Sample Interdisciplinary Data for Projects:

- Human interactions and preferences (e.g. from Google, Facebook, Netflix, Wikipedia, American Time Use Survey, Chinese Census Data), entrepreneurship (data on MBA programs, world bank data), activity generated data (e.g. web tracking, crowd sourcing, apps, Pandora, driving data).
- Biology - data on the numbers and types of fauna in a particular park over many years, human movement data, genomics, ecological forecasting, epidemiology, healthcare data, NIST data.
- Astronomy - Radio jet data, FITS image data
- Physics - protein folding, human movement data, statistical mechanics applies statistics to small particle theory
- Geoscience - ice core data for global temperatures, Paleobiology Database.
- Neuroscience - neuron firing rates as objects move in field of vision, connectomics, determining what traits are predicted by IQ, classifying personality types via relevant factors
- Chemistry - understanding reactions and energy via simulation, analyzing concentrations of a chemical in a sample
- Psychology - data on brain disorders and genetic correlations
- Political science - data on polarization in Congress; socially supplied data, data from religious leaders and congregations, bureau of labor statistics, voting patterns
- Economics - financial data, linear regression to relate economic quantities, confirmation of supply and demand in various sectors of the market, gender vs. pay, race vs. pay, education vs. crime, etc.
- Computer science - analysis of social networks and how information spreads, running times of randomized algorithms, numerous other examples
- Education and Sociology - general social survey
- Sports (sabremetrics in baseball, the role of data in football, soccer, and basketball)
- Computational linguistics (e.g. using data to translate dead languages, natural language processing)