IDS 702: Modeling and Representation of Data

Fall 2019

Duke University

Instructor: *Olanrewaju Michael Akande*

Email: *olanrewaju.akande@duke.edu*

Course Website: *TBD*

Lectures: *TBD*

Labs: *MAYBE?*

Office: *256 Gross Hall*

Office Hours: *TBD, 256 Gross Hall. Other hours available by appointment.*

Textbook: *TBD.*

Reading materials: *Available on course website. Sakai?*

Important Dates: *Friday, October 4, 7:30pm – Fall break begins*

Wednesday, October 9, 8:30am – Fall break ends

Tuesday, November 26, 10:30pm – Thanksgiving recess; Graduate classes end

---

Course Overview

Statistical models are necessary for analyzing the type of multivariate (often large) datasets that are usually encountered in data science and statistical science. This is a graduate level course that aims to provide students with the statistical data analysis tools needed to succeed as data scientists. In this course, you will learn several parametric modeling techniques such as generalized linear models, models for multilevel data and time series models. You will also learn to handle messy data, including data with missing or erroneous values, and data with outliers or non-standard distributions. You will be able to assess model fit, validate model assumptions and more generally, check whether proposed statistical models are appropriate for any given data. You will also learn causal inference under the potential outcomes framework. Should time permit, you may also learn nonparametric models such as classification and regression trees. Although this course emphasizes data analysis over rigorous mathematical theory, students who wish to explore the mathematical theory in more detail than what is covered in class are welcome to engage with and request further reading materials from the instructor outside of class.

Learning Objectives

By the end of this course, students should be able to

- Use the statistical methods and models covered in class to analyze real multivariate data that intersect with various fields.
- Assess the adequacy of statistical models to any given data and make a decision on what to do in cases when certain models are not appropriate for a given dataset.
- Cleanup and analyze messy datasets using approaches covered in class.
- Hone collaborative and presentations skills through the process of consistent team work on and class presentations of team projects.

Prerequisites

Students are expected to know all topics covered in the MIDS summer course review and boot camp. These include basic statistical inference including significance tests and confidence intervals, linear regression with one predictor, and exploratory data analysis methods. Students are also expected to be familiar with R. Due to space constraints, the course is open only to students in the MIDS program.

Course Format

- This is a very hands-on data modeling class. Each class will feature analysis on at least one dataset and students will be expected to analyze many datasets in class. Therefore, remember to bring your laptops to class.
- Lectures will include a combination of notes/slides, working through theory in class and some group work to understand applications. Notes will be available on the website by 11:59pm the day before each class and are provided in advance to allow you pay more attention in class. Make sure to read the notes before class and perhaps print them so you can focus better and take notes instead. You are responsible for all the material covered in class and assigned textbook readings. Ask questions in class, during office hours or send an e-mail, but do not wait until the last minute.
- There will be 15-min in-class quizzes Tuesdays and Thursdays.
- Homework assignments will be posted immediately after class on Fridays and will be due at the beginning of class the following Wednesday. The assignments are to help you develop a better

understanding of the material covered in class and prepare for exams, so take them seriously! You must show ALL work to receive credit. You are encouraged to work with each other on the homework problems, but you must turn in your own work. If you copy someone else's work, both parties will fail the assignment and be reported to the Office of Student Conduct. If you have any questions about what constitutes plagiarism, do not hesitate to ask.

- Lab assignments MUST be submitted by 11:59pm the same day. The objective of the lab is to give you hands-on experience with data analysis using modern statistical software. We will use a statistical analysis package called RStudio, which is a front end for the R statistical language.

Class Materials

Lecture notes, labs and other reading resources will be posted on the course website while homework assignments and practice questions will be posted on Sakai.

Grading

- There are no make-ups for assignments or projects except for medical or familial emergencies or for reasons approved by the instructor before the due date. See the instructor in advance of relevant due dates to discuss possible alternatives.
- Your final grade will be determined as follows:

Component	Percentage
Class Participation	5%
Lab Reports	10%
Quizzes	15%
Homework	20%
Midterm	20%
Final Exam	30%

- Grades may be curved at the end of the semester. Cumulative averages of 90% – 100% are guaranteed at least an A-, 80% – 89% at least a B-, and 70% – 79% at least a C-, however the exact ranges for letter grades will be determined after the final exam.

- There will be 8–10 quizzes, 5 homework assignments, and 8–10 labs. The two lowest quiz scores and the lowest homework score will be dropped (this should give you enough cover for genuine and unavoidable absences). There will be no labs on the sixth week to provide you with more time to prepare for the final exam.

Late Submission Policy

- You will lose 40% of points on each homework if you submit a day after it is due and 100% if you submit later than that.
- You will lose 50% of points on each lab if you submit a day after it is due and 100% if you submit later than that.

Course Schedule *

We will cover the topics below. We may spend different amounts of time on each topic, depending on the interests of students.

1. Introduction to course
2. Linear regression: methods and practice
   (a) Multiple regression
   (b) Model assessment and validation
   (c) Multicollinearity
   (d) Heteroscedasticity
3. Logistic regression: methods and practice
4. Multinomial logistic and Poisson regression
5. Introduction to multilevel models
   (a) Fixed effects vs random effects
   (b) Mixed effects models
6. Dealing with messy data: missing values, errors, and outliers

*This is a tentative outline and it will be updated as we proceed. See the course website for a detailed schedule.

7. Time series models

8. Methods for causal inference

   (a) Association vs. causation, confounding, and Simpson's paradox.

   (b) The potential outcome framework: potential outcomes, assignment mechanism and estimands.

   (c) Observational studies with ignorable assignment mechanism and propensity scores: assumptions of ignorability (unconfoundedness), matching, weighting, regression, and propensity scores methods.

9. Wrap up: Models versus algorithms

Week 1 (Chapters 1-2)

Interpretations and definition of probability, experiments and events, summary statistics and histograms, permutations and combinations, conditional probability, independent events, and Bayes' theorem.

Week 2 (Chapters 3-5)

Introduction to random variables, probability mass functions, cumulative distribution functions, discrete distributions, probability density functions, continuous distributions, marginal, joint and conditional distributions, expectations – mean, variance, covariance and correlation –, and introduction to some special distributions – Bernoulli, Binomial, hyper-geometric, Poisson, negative binomial, multinomial, gamma and normal distributions.

Week 3 (Chapters 6-7)

The law of large numbers, central limit theorem and continuity correction, Bayesian estimation and inference, prior and posterior distributions, conjugacy, maximum likelihood estimators and their properties, improving an estimator, sufficient statistics, distributions of linear combinations, and functions of random variables.

Week 4 (Chapters 8-9)

Sampling distribution of a statistic, confidence and credible intervals, interpreting confidence and credible intervals, some specific confidence intervals, unbiased estimators, the student-t and Chi-square distributions, simple hypothesis testing, type I and II errors, two-sided hypothesis testing, power calculations, and introduction to Bayesian hypothesis testing.

Week 5 (Chapters 10-11)

Tests of independence, goodness of fit tests, contingency tables, Simpson's paradox, the method of least squares and simple linear regression, introduction to multiple, nonlinear and nonparametric regression, model validation and assessment tools, and one-way analysis of variance.

Week 6 (Chapters 11-12)

Two-way analysis of variance, simulation, bootstrap; REVISION.

Academic Integrity

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Cheating on exams or plagiarism on homework assignments, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved. Additionally, there may be penalties to your final class grade along with being reported to the Undergraduate Conduct Board. Please review the academic dishonesty policies at `https://studentaffairs.duke.edu/conduct/z-policies/academic-dishonesty`.

Disability Statement

Students with disabilities who believe that they may need accommodations in the class are encouraged to contact the Student Disabilities Access Office at 919.668.1267 or disabilities@aas.duke.edu as soon as possible to better ensure that such accommodations are implemented in a timely fashion.

Other Information

For every lecture, you will need a simple calculator for quizzes, exams and homework. Graphical capability is not required, and questions are worded so that advanced calculators confer no advantage. I will provide

you any other material needed for the quizzes and exam. Again, do not hesitate to come to my office during office hours or by appointment to discuss a homework problem or any aspect of the course, and working in groups is highly recommended. Questions related to course assignments and honesty policy should be directed to me. DO NOT search for direct answers to homework questions online; ask me instead.