IDS 702: Data Modeling and Representation

Fall 2024 Duke University

Course Overview

Meeting times: Tuesdays and Thursdays 3:05-4:20 PM, Gross Hall 330

Course Objectives

Developing an understanding of statistical modeling is a key component of becoming a data scientist. Statistical models are used to answer research questions and obtain meaningful insights from many kinds of data.

Broadly, this course will cover the following topics:

1. Statistics Fundamentals
2. Linear Regression
3. Generalized Linear Models

But here in the MIDS program, understanding the content is only the beginning. Successful data scientists are **critical thinkers**, **problem solvers**, **effective communicators**, and **enthusiastic collaborators**. With that in mind, this course aims to meet five key learning objectives:

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

1. Fit and interpret statistical models, including linear and generalized linear models.
2. Connect statistical modeling concepts to underlying statistics fundamentals including probability distributions and estimation.
3. Map a research question and dataset to the appropriate statistical model.
4. Make careful and critical decisions about model building and consider real-world implications.

5. Communicate (through written and oral communication) model results to a broad audience.

Course Materials

Textbooks

I will assign readings from a few different textbooks throughout the semester. All of these books are available for free in electronic format either online or through the Duke Library.

Statistics for Data Scientists: An Introduction to Probability, Statistics, and Data Analysis by Kaptein, M. and van den Heuvel, E. (Available through the Duke Library)

An Introduction to Statistical Learning with Applications in R, 2nd edition by James, G., Witten, D., Hastie, T., and Tibshirani, R. (Available online)

Statistical Foundations, Reasoning and Inference For Statistics and Data Science by Kauermann, G., Küchenhoff, H., and Heumann, C. (Available through the Duke Library)

Gradescope

You will submit assignments through Gradescope. You can use this code to access the course: **J74D7E**

Slack

Be sure to join the ids702-fall24 channel in the MIDS Slack workspace. I will post announcements on this channel.

A note on communication: If you would like to contact me for any reason, you can reach out on either Slack **or** via email. If I do not respond within 24 hours during weekdays, you are welcome to follow up.

Course Components

Class preparation

Before each class meeting, you will be required to engage with prep materials (reading or video). The prep materials will primarily cover theoretical modeling concepts. Class meetings will then focus on implementation in R and application exercises. Preparation assignments will be posted on the course website.

To ensure that you engage with the class preparation, we will have random comprehension quizzes throughout the semester. These quizzes will take place at the beginning of class and will count as part of your participation grade (cumulative score >80% will earn full credit).

The preparation course component connects to the first two learning objectives.

Application exercises

During class meetings, you will complete application exercises. These exercises will focus on deepening your understanding of the theoretical concepts and help you to apply the concepts in R. You will work on these exercises in groups and submit it at the end of the class session. They will be graded on the following scale:

0: Exercise was not submitted

1: Exercise was submitted but does not present a good faith effort

2: Exercise was submitted and presents a good faith effort

A "good faith effort" means that the exercise may not be completely correct, or may be missing a couple of components, but it is clear that students worked diligently and engaged in the learning process.

The preparation course component connects to the first, second, and fourth learning objectives.

Homework assignments

You will have five homework assignments to complete during the semester; they will each cover roughly two weeks of material. The structure of the assignments will evolve with the course material, starting with a problem set structure and moving toward written analysis assignments.

You are encouraged to talk to each other about general concepts, and to the instructor/TAs. However, the write-ups, solutions, and code MUST be entirely your own work. The assignments must be typed up using Quarto and submitted on Gradescope. Note that you will not be able to make online submissions after the due date, so be sure to submit before the Gradescope-specified deadline.

Homework assignments connect to all five learning objectives.

Statistics reflections

You will be responsible for four statistics reflections throughout the semester. I have assigned six pieces that cover various topics related to the interaction between data science and society. You are to write a written reflection about the material that you select. Questions are provided as prompts, but you are not required to answer them in your reflection. Grades will be based on completion and thoughtful engagement.

The chosen articles/videos address topics that may be sensitive and/or uncomfortable, including racism, eugenics, and gender identity. It is crucial that you engage in the reflections thoughtfully and respectfully. I seek to create a classroom environment that not only acknowledges diversity in all forms, but celebrates it. In that endeavor, I would be remiss not to acknowledge the discriminatory ways statistical science has been used both historically and currently. In having these important discussions, I want you to critically examine and appreciate the power (good and bad) of statistics as you begin your career as a data scientist. More information can be found on the statistics reflections page on the course website.

Statistics reflections connect to the fourth course learning objective.

Midterm exam

A midterm exam will be held on Tuesday, October 22, in class, and cover the first two units: statistics fundamentals and linear regression. The exam will assess conceptual understanding, not R programming.

If you benefit from the use of testing accommodations, such as an isolated testing area or extra time, please connect with the Student Disability Access Office as soon as possible.

The exam connects to the first and second learning objectives.

Project

You will work with a team to apply the knowledge and skills learned throughout this course and analyze a dataset that interests you. The project should be an in-depth statistical analysis of a particular research question. Your team will select the dataset. Teams will be assigned. More detailed information will be available on the course website later in the semester. The project will have the following components:

- Proposal (team)
- Written report (team)
- Recorded elevator pitch (individual)
- Team member evaluation

The project connects to all four course learning objectives.

Grade Calculation

Component	Percentage
Participation (prep quizzes and application exercises)	10%
Statistics reflections	10%
Homework assignments	40%
Midterm exam	15%
Project	25%

Letter grade scales may be adjusted at the end of the semester. Cumulative averages $\geq$ 90% are guaranteed at least an A-, cumulative averages $\geq$ 80% are guaranteed at least a B-, and cumulative averages $\geq$ 70% are guaranteed at least a C-

Regrade requests can be made on Gradescope within 24 hours of the assignment's grade release. Regrade requests for final project reports/presentations must be made within 12 hours of grade release.

There are no make-ups for any graded work except for cases of medical/personal/familial emergencies. If you are facing extenuating circumstances during the semester, please don't hesitate to reach out to the instructor.

Course policies

Late submissions

You (or your team when applicable) will lose 50% of the total points on each assignment if you submit within the first 24 hours after it is due. You will lose 100% of the total points if you submit later than that without explicit approval from the instructor.

You can request a single-use, no-questions-asked 24-hour extension for one assignment (either homework assignment or statistics reflection) during the semester. To use your extension, email the instructor and cc all TAs with the subject line "IDS 702 Single-use Extension." No explanation is needed as to why you are requesting the extension. The extension cannot be applied to the exam, project deliverables, or prep material quizzes.

Academic integrity

As a student in this course, you have agreed to uphold the Duke Community Standard as well as the practices specific to this course. This means that the work you submit is your own, even if you discuss assignments with your classmates. Additionally, when consulting resources (books, internet articles including stackexchange, chatgpt), you must cite them. If you have any questions about what should be cited in your work, please reach out to the instructor.

Please read Nick Eubank's ChatGPT advice linked here and consider the downside of overreliance on LLMs during the learning process.

Inclusive community

It is my intent that students from all diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that the students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity and in alignment with Duke's commitment to diversity and inclusion. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.

Furthermore, I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities. To help accomplish this:

- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. If you prefer to speak with someone outside of the course, I encourage you to speak with MIDS administrators.
- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please let me or a member of the teaching team know.

Resources

- Duke Counseling & Psychological Services (CAPS) helps Duke Students enhance strengths and develop abilities to successfully live, grow and learn in their personal and academic lives. CAPS offers many services to Duke students, including brief individual and group counseling, couples counseling and more. CAPS staff also provides outreach to student groups, particularly programs supportive of at-risk populations, on a wide range of issues impacting them in various aspects of campus life. CAPS provides services to students via Telehealth. To initiate services, you can contact their front desk at 919-660-1000.

- If there is any portion of the course that is not accessible to you due to challenges with technology or the course format, please let me know so we can make appropriate accommodations. The Student Disability Access Office (SDAO) is available to ensure that students are able to engage with their courses and related assignments.
- The Academic resource center provides learning resources to help you maximize your academic capabilities.

Schedule

Individual lecture topics are subject to change. The midterm date will not change.

Week of	Tuesday content	Thursday content	Assig
Aug 26	Study Design	Intro Probability	
Sept 2	Random variables & Distributions	Sampling Distributions & Central Limit Theorem	Stati
Sept 9	Estimation	Confidence Intervals/Bootstrap	Hom
Sept 16	Inference I	Inference II	Stati
Sept 23	Intro to Regression	Multiple Linear Regression	Hom
Sept 30	Categorical Predictors, Interaction terms	Assessing Assumptions	Stati
Oct 7	Influential points & Multicollinearity	Model Selection	Hom
Oct 14	Fall break - no class	Exam review	
Oct 21	Midterm Exam	Intro to GLMs	
Oct 28	Logistic Regression	Assessing Logistic Models	Stati
Nov 4	Multinomial regression	Ordinal regression	Hom
Nov 11	Poisson regression	Survival Analysis	Proj
Nov 18	Missing Data	Missing Data	Hom
Nov 25	Project work day		
			Dec