IDS 702: Modeling and Representation of Data

Fall 2021
Duke University

Instructor: Olanrewaju Michael Akande, Ph.D.

Email: ✉ olanrewaju.akande@duke.edu

Office: **223 Gross Hall**

Office Hours: **Wednesdays and Fridays (9am – 10am)**.
Zoom meeting ID: **See Sakai**.

Course Page: https://ids702-f21.olanrewajuakande.com

Meeting Times: **Tuesdays and Thursdays (10:15 – 11:30am), 270 Gross Hall.**

Teaching Assistants: Jiaman Betty Wu.
Tuesdays (9am – 10am) and Thursdays (5:30pm – 6:30pm).
Zoom meeting ID: **See Sakai**.

Xinyi (Iris) Pan.
Mondays (8am – 9am) and Fridays (5pm – 6pm).
Zoom meeting ID: **See Sakai**.

Recommended Textbooks: *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Gelman A., and Hill, J.
An Introduction to Statistical Learning with Applications in R by James, G., Witten, D., Hastie, T., and Tibshirani, R.
While recommended, these books are not compulsory. That said, they are both really great books, so be sure to get a copy if you can (free pdf of ISL is available via the link)!

Optional Textbooks: *An Introduction to Categorical Data Analysis, Second Edition* by Alan Agresti.
I will assign some optional readings for a few topics from this book. You can download pdf versions of individual chapters, via Duke library using the link.

Important Dates:

• Monday, August 23	Fall classes begin
• Friday, September 3	Drop/Add ends
• Monday, September 6	Labor day. Classes in session
• Sunday, October 3	Team project I reports due
• Sunday, October 24	Team project II reports due
• Wednesday, October 27	Final project proposal due
• Monday, November 22	Upload final project presentations
• Tuesday, November 23	End of semester
• Sunday, December 12	Final project reports due (tentative)

1 Course Overview

Statistical models are necessary for analyzing the multivariate (often large) datasets that are usually encountered in data science and statistical science. This graduate level course, a core part of Duke's Master in Interdisciplinary Data Science (MIDS) program, aims to provide students with the statistical data analysis tools needed to succeed as data scientists.

In this course, you will learn the general work flow for building statistical models and using them to answer inferential questions. You will learn several parametric modeling techniques such as linear regression, generalized linear models, models for multilevel data and basic time series models. You will also learn to handle messy data, including data with missing values, assess model fit, and validate model assumptions and more generally, check whether proposed statistical models are appropriate for any given data. You will also learn a bit of causal inference under the potential outcomes framework and should time permit, a bit of 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.

Finally, this course is designed primarily for students in the MIDS program. Enrollment for non-MIDSters is subject to numbers and permission will be granted on a case-by-case basis.

2 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.

3 Course Info

3.1 Course Format

This course is designed to be primarily synchronous. However, there will also be some asynchronous activities. Students will be required to do pre-assigned readings, go through lecture slides, watch some pre-recorded lecture videos, and take the quizzes embedded in the videos, all before each synchronous meeting times. The meeting times are therefore primarily reserved for in-class activities, discussions and Q&A sessions.

3.2 Playposit

To gain access to the pre-recorded lecture videos, you will have to create a Playposit account. There are participation quizzes embedded within the videos. These quizzes make up a part of your final grade (see: course policies) so take them seriously. To join the class on Playposit, you need to create a new account as a student here, then use the class link here to join the class. While you need not create an account with your Duke email, I strongly suggest you do.

4 Prerequisites

Students are expected to know all topics covered in the MIDS summer course review and boot camp. These include basic probability and statistical inference, including random variables, probability distributions, central limit theorem, hypothesis testing, confidence intervals, linear regression with one predictor, and exploratory data analysis methods. Students are also expected to be familiar with R/RStudio and are encouraged to have learned LaTeX or a Markdown language by the end of the course. MIDS students automatically satisfy these requirements. If you are not a MIDS student, email the instructor to ascertain that you have taken courses that cover these topics.

5 Team Work

Note that this course, as is the case with most core courses within the MIDS program, emphasizes the ability to work in teams so that students can learn team productivity and performance. Each student must therefore be ready to contribute to their team's success. MIDS students will work in the same teams they have been assigned to for the fall semester. If you are not a MIDS student, you will be assigned to a group with other non-MIDS students, and by enrolling in this course, you are agreeing to being held to the same standard as MIDS students. Consequently, you are expected to be fully committed to team excellence, performance, and productivity.

6 Class Materials

Lecture notes and slides, links to the videos and other reading resources will be posted on the course website. We will only loosely follow the textbooks.

7 Graded Work

Graded work for the course will consist of participation quizzes, data analysis assignments, team projects, and a final project. Regrade requests for data analysis assignments and team projects must be done via Gradescope AT MOST **24 hours** after grades are released! Regrade requests for the final project must be done via Gradescope AT MOST **12 hours** after grades are released!

- There are no make-ups for any of the graded work except for medical or familial emergencies or for reasons approved by the instructor BEFORE the due date. Contact the instructor in advance of relevant due dates to discuss possible alternatives.

- 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 at the end of the course.

- There is no final exam. Students' final grades will be determined as follows:

Component	Percentage
Data Analysis Assignments	30%
Final Project	25%
Team Project I	17.5%
Team Project II	17.5%
Participation	10%

8 Descriptions of Graded Work

8.1 Data Analysis Assignments

Data analysis assignments will be posted on the course website. The assignments include questions that ask students to apply the statistical modeling skills discussed during the semester, as well as questions on the computational aspects of the methods. Students must turn in these assignments on the due date.

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

Solutions to the assignments will be curated from student solutions with proper attribution. Every week the TAs will select one or two representative solutions for the assigned problems with each solution being attributed to the student who wrote it. **If you would like to OPT OUT of having your solutions used for as a representative solution, let the Instructor and TAs know in advance.**

Finally, students may be asked to work in pairs for one or two of the data analysis assignments when possible. When that is the case, each pair need only submit one solution per assignment.

8.2 Final Project

For the final project, you will apply the knowledge and skills learned throughout this course to analyze a dataset that interests you, subject to the instructor's approval. The project should be an in-depth statistical analysis of a question that interests you. It is quite common for this final project to be based on your research interests, or topics/questions from one of your other courses. Just about every discipline has questions that are amenable to statistical analyses, including economics, engineering, environmental studies, history, the natural sciences, psychology, and even sports, so there are many options to choose from. The data should comprise several variables amenable to statistical analyses via modeling. Students can bring in their own research data sets, or they can ask the instructor for assistance with identifying appropriate data. You will be expected to present the results of your analysis. Detailed instructions will be made available later.

8.3 Team Projects

For the team projects, students will work in teams to analyze data selected by the instructor. Each team will be expected to write a report with their data analysis findings. Students may also be given the opportunity to present their results in class. Detailed instructions will be made available later.

8.4 Participation

Each student will be assigned a participation grade based on their level of participation throughout the semester. Participation will be assessed based on performance on PlayPosit and in-class quizzes, engagement during live meeting sessions and breakout rooms, and generally how each students engages with other students on Ed Discussion, especially regarding feedback on the project presentations.

9 Late Submission Policy

You (or your team when applicable) will lose 50% of the total points on each data analysis assignment, each team project, and the final project, 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.

10 Tentative Course Schedule

We will cover the topics below. We may spend different amounts of time on each topic, depending on the interests of students. For a detailed and updated outline, check on the updated course schedule on the course page regularly.

- Introduction to course
- Linear regression
  - Introduction to multiple linear regression
  - Inference and prediction
  - Model assessment and diagnostics
  - Transformations and multicollinearity
  - Model building and selection
- Logistic regression
  - Introduction
  - Interpretation of coefficients
  - Inference vs prediction
  - Model assessment and validation
- Other generalized linear models
  - Multinomial logistic regression
  - Proportional odds model
  - Poisson regression
  - Probit regression
- Introduction to multilevel models
  - Fixed vs random effects
  - Multilevel linear models
  - Multilevel logistic regression
- Dealing with messy data
  - Missing values, errors, and outliers
  - Single imputation methods
  - Multiple imputation
- Methods for causal inference
  - Introduction, association vs. causation, and confounding variables
  - Observational studies: regression, stratification and matching
  - Observational studies: propensity scores methods
- Basic time series models
  - AR and MA models
  - ARMA and ARIMA models
- Random number generation, bootstrap and Monte Carlo
- Introduction to polynomial regression, local regression, and tree-based methods
- Wrap up and final projects

11 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 or plagiarism on any graded assessments, 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 Office of Student Conduct. Review the academic dishonesty policies at https://studentaffairs.duke.edu/conduct/z-policies/academic-dishonesty.

12 Diversity & Inclusiveness:

This course is designed so that students from all backgrounds and perspectives all feel welcome both in and out of class. Please feel free to talk to me (in person or via email) if you do not feel well-served by any aspect of this class, or if some aspect of class is not welcoming or accessible to you. My goal is for you to succeed in this course, therefore, let me know immediately if you feel you are struggling with any part of the course more than you know how to manage. Doing so will not affect your grades, but it will allow me to provide the resources to help you succeed in the course.

13 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.

14 Other Information

It can be a lot more pleasant oftentimes to get one-on-one answers and help. Make use of the teaching team's office hours, we're here to help! Do not hesitate to talk to me during office hours or by appointment to discuss a problem set or any aspect of the course. Questions related to course assignments and honesty policy should be directed to me. When the teaching team has announcements for you we will send an email to your Duke email address. Be sure to check your email daily.

If you have any concerns, issues or challenges, let the instructor know as soon as possible. Also, all students are strongly encouraged to rely on Ed Discussion, for interacting among yourself and asking other students questions. You can also ask the instructor or the TAs questions on there and we will try to respond as soon as possible. If you experience any technical issues with joining or using Ed Discussion, let the instructor know.

15 Professionalism

Try as much as possible to refrain from texting or using your computer for anything other than coursework while watching the lecture videos or while in class. Again, the more engaged you are, the quicker you will be able to get through the materials. You are responsible for everything covered in the lecture videos, lecture notes/slides, and in the assigned readings.