

IDS 702: MODELING AND REPRESENTATION OF DATA

FALL 2019
DUKE UNIVERSITY

INSTRUCTOR:	OLANREWAJU M. AKANDE, PH.D.
EMAIL:	✉ olanrewaju.akande@duke.edu
OFFICE:	256 Gross Hall
OFFICE HOURS:	Mon/Tue 3:00 - 4:00pm, 256 Gross Hall. Other hours available by appointment.
TEACHING ASSISTANTS:	Azucena Morales . Office hours: Thur 2:30 - 4:30pm. Chenxi Wu. Office hours: Wed 3:00 - 5:00pm. <i>All TA office hours will hold in Gross Hall.</i>
LECTURES:	Tue/Thur, 10:05 - 11:20am, 270 Gross Hall.
LABS:	Every other Fri (starting Sep 6), 10:00 - 11:45am, 270 Gross Hall.
COURSE PAGE:	https://akandelanre.github.io/IDS702_F19/
RECOMMENDED TEXTBOOKS:	Data Analysis Using Regression and Multilevel/Hierarchical Models by Gelman A., and Hill, J.
OPTIONAL TEXTBOOKS:	An Introduction to Statistical Learning with Applications in R by James, G., Witten, D., Hastie, T., and Tibshirani, R. (Free pdf available online via the link.)
IMPORTANT DATES:	Fri, October 4, 7:30pm – Fall break begins Wed, October 9, 8:30am – Fall break ends Mon, November 4, 11:59pm – Final project proposal due Tue, November 26, 10:30pm – Thanksgiving/graduate classes end Tue, December 3 and Thur, December 5 – Final project presentations

1 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, within the curriculum for Duke's Master in Interdisciplinary Data Science (MIDS) program, that 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 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, we may also briefly cover 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.

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 Prerequisites

Students are expected to know all topics covered in the MIDS summer course review and boot camp. These include basic statistical inference including hypothesis testing, 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.

4 Class Materials

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

5 Graded Work

Graded work for the course will consist of methods and data analysis assignments, team projects, and a final project.

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

Component	Percentage
Methods and Data Analysis Assignments	30%
Final Project	30%
Team Project 1	15%
Team Project 2	15%
Lab Assignments	10%

- ▀ There are no make-ups for assignments or the 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.
- ▀ 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.

6 Descriptions of graded work

6.1 Methods and Data Analysis Assignments

Methods and Data Analysis assignments are posted on the IDS 702 course website. Students turn in these assignments at the beginning of class on the due date. Students are permitted to work with others on the assignments, but each person must write up and turn in her or his own answers. The Methods and Data Analysis assignments include questions on the computational and the mathematical aspects of the methods that underpin the statistical models we learn during the semester, and questions that ask students to apply the modeling skills discussed during the semester. The assignments must be typed up using R Markdown, L^AT_EX or another word processor, and submitted on Sakai 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 Sakai-specified deadline.

6.2 Lab Assignments

The objective of the lab assignments is to give you more hands-on experience with data analysis using R. The labs times also gives you an additional platform to ask for help for your team and individual projects. Lab attendance is not mandatory on days when team presentations will not hold, however, each lab assignment should be submitted in timely fashion on the due date.

6.3 Team Projects

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

6.4 Final Project

For the final project, students analyze a data-based research question of their choosing, subject to the instructor's approval. 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. Students present their results in a poster session. Detailed instructions will be made available later on the website.

7 Late Submission Policy

- You will lose 50% of the total points on each homework if you submit within the first 24 hours after it is due, and 100% of the total points if you submit later than that.
- You will lose 40% of the total points on each lab if you submit within the first 24 hours after it is due, and 100% of the total points if you submit later than that.

8 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
- ▀ Multiple predictors

- Inference and prediction
- Model assessment, diagnostics and validation
- Transformations, multicollinearity and heteroscedasticity
- Model building and selection
- ▶ Logistic regression
 - Interpreting logistic regression coefficients
 - Inference vs prediction
 - Model assessment and validation
- ▶ Generalized linear models
 - Multinomial logistic
 - Poisson regression
 - Probit regression
- ▶ Introduction to multilevel models
 - Fixed effects 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
 - Association vs. causation, confounding, and Simpson's paradox.
 - The potential outcome framework: potential outcomes, assignment mechanism and estimands.
 - Observational studies with ignorable assignment mechanism and propensity scores: assumptions of ignorability (unconfoundedness), matching, weighting, regression, and propensity scores methods.
- ▶ Time series models
- ▶ Wrap up (would cover introduction to nonparametric methods should time permit).

9 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>.

10 Diversity & Inclusiveness:

In line with the MIDS culture, 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, please 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.

11 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.

12 Other Information

It can be a lot more pleasant oftentimes to get in-person answers and help. Make use of the teaching team's office hours, we're here to help! 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. 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. Please make sure to check your email daily.

13 Professionalism

Please refrain from texting or using your computer for anything other than coursework during class.