BMI 510: Biostatistics for Machine Learning

Updated 2025-01-15

Instructors

J. Lucas McKay, Ph.D., M.S.C.R.
lucas@dbmi.emory.edu

Selen Bozkurt, Ph.D. M.S.
selen.bozkurt@dbmi.emory.edu

Dates

2025-01-15 through 2025-04-28 (Final Package due 2025-05-08)

Time

M 2:30 pm-3:45 pm (Lecture)
W 2:30 pm-3:45 pm (Lab)
Th 12:00 pm-12:50 pm (Journal Club)

Location

Woodruff Memorial Bldg. 4004

TA

Masoud Nateghi
bmi510@dbmi.emory.edu

Prerequisites

Some matrix algebra, any interpreted or compiled computing language.

Computing

We will use R/RStudio, both of which are available as free, open-source software. A laptop computer is necessary for some in-class exercises.

Accommodations

As the instructors of this course, we endeavor to provide an inclusive learning environment. We want every student to succeed. The Department of Accessibility Services (DAS) works with students who have disabilities to provide reasonable accommodations. It is your responsibility to request accommodations. In order to receive consideration for reasonable accommodations, you must register with the DAS at https://accessibility.emory.edu/students/. Accommodations cannot be retroactively applied, so you need to contact DAS as early as possible and contact us as early as possible in the semester to discuss the plan for implementation of your accommodations. For additional information about accessibility and accommodations, please contact the DAS at (404) 727-9877 or accessibility@emory.edu.

Support

As a student, you may find that personal and academic stressors in your life, including those related to illness, economic instability, and/or racial injustice, are creating barriers to learning this semester. Many students face personal and environmental challenges that can interfere with their academic success and overall wellbeing. If you are struggling with this class, please visit me during office hours or contact me via email. If you are feeling overwhelmed and think you might benefit from additional support, please know that there are people who care and offices to support you at Emory. These services – including confidential resources – are provided by staff who are respectful of students' diverse backgrounds. For an extensive list of well-being resources on campus, please go to: http://campuslife.emory.edu/support/index.html. Keep in mind that Emory offers free, 24/7 emotional, mental health, and medical support resources via TimelyCare: https://timelycare.com/emory.

Academic Integrity

You are expected to uphold and cooperate in maintaining academic integrity as a member of the Laney Graduate School. By taking this course, you affirm your commitment to the Laney Graduate School Honor Code, which you can find in the Laney Graduate School Handbook. You should ensure that you are familiar with the rights and responsibilities of members of our academic community and with policies that apply to students as members of our academic community. Any individual, when they suspect that an offense of academic misconduct has occurred, shall report this suspected breach to the appropriate Director of Graduate Studies, Program Director, or Dean of the Laney Graduate School. If an allegation is reported to a Director of Graduate Studies or a Program Director, they are in turn required to report the allegation to the Dean of Laney Graduate School.

Collaboration

Homework. You are encouraged to discuss homework problems with each other at a conceptual / pseudocode level. Like most CS classes, you are not allowed to directly share code with other students. Exams. *Exams.* Exams are take-home, open-book, similar to long homework exercises. You may not discuss exams with other students. *Final project.* The final project is a complete, documented R package comprising functions created in the class as well as some new ones. You may not discuss the final project with other students.

LLMs

You may use LLMs like CoPilot and ChatGPT for all exercises. However, you are responsible for the functionality of the code.

Summary

This course presents an accelerated introduction to the concepts and methods of biostatistical data analysis suitable for applying machine learning approaches on clinical data. Topics include exploratory data analysis with grammar-based data visualization (e.g., *ggplot2/seaborn*); dimensionality reduction, measures of model fit, descriptive statistics and confidence intervals for categorical, ordinal, and continuous variables with normal and non-normal distributions; measures of association; statistical power; one- and two-sample hypothesis tests; ANOVA and hierarchical linear models.

Office hours

F 12:00 pm-12:50 pm (please request via email)

Course objectives

By completion of the course, students will be able to choose and implement appropriate statistical analyses for a variety of data types; generate descriptive statistics for clinical data; conduct multivariate linear and logistic regression analyses; and describe and interpret their analyses. Students will be able to manage data and implement statistical tests in modern statistical software. One hour each week will be spent on critical analyses of bias and analytic methods in published primary and secondary literature.

Evaluation

Weekly homework	30 %
Journal club presentation	10 %
Quiz 1	20 %
Quiz 2	20 %
Final software package	20 %

Grading

95+	A	90 – 94	A-
85 – 89	B+	80 – 84	B
75 – 79	B-	65 – 74	C
<65	F		

Late penalties

2% of the total points for each late homework will be deducted for each hour it is late. No late midterm or final projects will be accepted.

Decorum

Students are expected to keep their cameras on and to engage during all class sessions if remote. Failure to do so may result in substantial penalties to the final grade. There is no need to keep your camera on if you are in the classroom with the slides up/similar.

Resources

Slides and other resources will be posted at https://jlucasmckay.bmi.emory.edu/global/bmi510

Optional Texts

Mathematical statistics and data analysis (Rice)
An Introduction to Statistical Learning with Applications in R (James)
Generalized linear models with examples in R (Dunn)

Course Design

BMI 510 is an accelerated course designed to get students a head start on data visualization, common analytic scenarios, and get them aware of and able to articulate the impacts that (inevitably) biased data may have on their work. The class meetings are of three main types:

1. Mondays are traditional lectures, although there may be some interactive computer exercises. These meetings will be semi-in-person; room 4004 is available to meet. Dr. McKay will give the lecture from his office a few doors over if COVID control or other reasons are necessary.
2. Wednesdays are “Lab” sessions, the majority of which will include a lecture component at the beginning followed in some cases by an interactive component.
3. Thursdays are “Journal Club” sessions. These will feature brief slide presentations followed by discussion sessions on articles related to bias in AI systems, or in human systems that are anticipated to be automated soon.

Zoom Meeting Information

There are two separate zoom meetings due to the separate start times.

Monday and Wednesday:

Join Zoom Meeting
https://zoom.us/j/95981972226

Meeting ID: 959 8197 2226

Thursdays:

Join Zoom Meeting
https://zoom.us/j/91769146332

Meeting ID: 917 6914 6332

Course Calendar

(subject to revision)

Day	Date	Topic	Instructor	Assigned	Due
W	1/15/2025	R/RStudio installation and housekeeping	M/B	HW0	
Th	1/16/2025	Bias	M		
M	1/20/2025	No Class - King Day			
W	1/22/2025	Introduction to Course	M/B	HW1	HW0
Th	1/23/2025	Journal Club: Fairness in AI	M		
M	1/27/2025	Probability	B		
W	1/29/2025	Random variables	M	HW2	HW1
Th	1/30/2025	Journal Club	S		
M	2/3/2025	Useful distributions	B		
W	2/5/2025	Data wrangling	M	HW3	HW2
Th	2/6/2025	Journal Club	S		
M	2/10/2025	One- and two-sample tests	B		
W	2/12/2025	Tests of proportions	B	HW4	HW3
Th	2/13/2025	Journal Club	S		
M	2/17/2025	Power analysis	M		
W	2/19/2025	Multiple comparisons	M	HW5	HW4
Th	2/20/2025	Journal Club	S		
M	2/24/2025	Confidence intervals	M		
W	2/26/2025	Fitting and simulating distributions	M	HW6	HW5
Th	2/27/2025	Journal Club	S		
M	3/3/2025	Analysis of variance	M		
W	3/5/2025	Table one	M	MT1	HW6
Th	3/6/2025	Journal Club	S		
M	3/10/2025	No class - Spring Break			
W	3/12/2025	No class - Spring Break			
Th	3/13/2025	No class - Spring Break			
M	3/17/2025	Classifiers	M		
W	3/19/2025	Assessing model performance	M	HW7	MT1
Th	3/20/2025	Journal Club	S		
M	3/24/2025	Linear models I - Introduction	B		
W	3/26/2025	Linear models II - Hypothesis	B	HW8	HW7

		testing			
Th	3/27/2025	Journal Club	S		
M	3/31/2025	Linear models III - Multiple regression	M		
W	4/2/2025	Linear models IV - Variable selection, regularized regression	M	HW9	HW8
Th	4/3/2025	Journal Club	S		
M	4/7/2025	Logistic regression I	B		
W	4/9/2025	Logistic regression II	B	HW10	HW9
Th	4/10/2025	Journal Club	S		
M	4/14/2025	Linear mixed models – fixed / random effects	M		
W	4/16/2025	Deviance	M	MT2	HW10
Th	4/17/2025	Journal Club: Ethics	B		
M	4/21/2025	Fairness metrics	B		
W	4/23/2025	R packages	M	Final Package	
Th	4/24/2025	Journal Club: Stochastic Parrots	M/B		
M	4/28/2025	Wrap-up	M		
W	4/30/2025	No class - Finals			MT2
Th	5/1/2025	No class - Finals			
M	5/5/2025	No class - Finals			
W	5/7/2025	No class - Finals			
Th	5/8/2025	Final Package Due			Final Package