**BMI 510: Biostatistics for Machine Learning**

Instructor J. Lucas McKay, Ph.D., M.S.C.R.

 lucas@dbmi.emory.edu

 bmi510@dbmi.emory.edu

Dates 2024-01-17 through 2024-04-29 (Final Package due 2024-05-06)

Time M 1:00 pm-2:15 pm (Lecture)

W 1:00 pm-2:15 pm (Lab)

 Th 12:00 pm-12:50 pm (Journal Club)

Location Woodruff Memorial Bldg. 4004

TA Sean Jordan

 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 Please discuss any necessary accommodations with Dr. McKay.

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 mid-term or final projects will be accepted.

Collaboration Similar to other CS classes, you may not share code with your classmates under any circumstances. However, **you may discuss general statistical concepts as well as the overall approach required for particular homework problems at a “pseudocode” level**, if needed.

**Midterms are to be completed on your own**; no discussion whatsoever is permitted until the solutions are revealed.

**The final package is to be completed on your own**; no discussion whatsoever is permitted until the solutions are revealed.

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 necessitate.
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/96540373027?pwd=VWUyRFgxaFA2Uzc4bm9pMlpBTWtTdz09](https://www.google.com/url?q=https://zoom.us/j/96540373027?pwd%3DVWUyRFgxaFA2Uzc4bm9pMlpBTWtTdz09&sa=D&source=calendar&ust=1705940789500909&usg=AOvVaw3YlRfkhPB3u1VgYbItcxEJ)

Meeting ID: 965 4037 3027

Passcode: 080264

***Thursdays:***

Join Zoom Meeting [https://zoom.us/j/96279354065?pwd=cThpMUpkWlNUSG8rN2tKVk1RZ092QT09](https://www.google.com/url?q=https://zoom.us/j/96279354065?pwd%3DcThpMUpkWlNUSG8rN2tKVk1RZ092QT09&sa=D&source=calendar&ust=1705940980722093&usg=AOvVaw2aymLMx4OV-t5h-jkMDxUV)

Meeting ID: 962 7935 4065

Passcode: 872398

**Course Calendar**

*(subject to revision)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Day** | **Topic** | **Due** | **Assigned** |
| 2024-01-17 | W | R/RStudio installation and housekeeping |  | HW-00 |
| 2024-01-18 | Th | Bias |  |  |
| 2024-01-22 | M | Introduction to Course | HW-00 | HW-01 |
| 2024-01-24 | W | Introduction to R / data visualization |  |  |
| 2024-01-25 | Th | Journal Club: Fairness in AI |  |  |
| 2024-01-29 | M | Probability | HW-01 | HW-02 |
| 2024-01-31 | W | Random variables |  |  |
| 2024-02-01 | Th | Journal Club |  |  |
| 2024-02-05 | M | Useful distributions | HW-02 | HW-03 |
| 2024-02-07 | W | Data wrangling |  |  |
| 2024-02-08 | Th | Journal Club |  |  |
| 2024-02-12 | M | One- and two-sample tests | HW-03 | HW-04 |
| 2024-02-14 | W | Tests of proportions |  |  |
| 2024-02-15 | Th | Journal Club |  |  |
| 2024-02-19 | M | Power analysis | HW-04 | HW-05 |
| 2024-02-21 | W | Multiple comparisons |  |  |
| 2024-02-22 | Th | Journal Club |  |  |
| 2024-02-26 | M | Confidence intervals | HW-05 | HW-06 |
| 2024-02-28 | W | Fitting and simulating distributions |  |  |
| 2024-02-29 | Th | Journal Club |  |  |
| 2024-03-04 | M | Analysis of variance | HW-06 | MT-01 |
| 2024-03-06 | W | Table one |  |  |
| 2024-03-07 | Th | Journal Club |  |  |
| 2024-03-11 | M | No class - Spring Break |  |  |
| 2024-03-13 | W | No class - Spring Break |  |  |
| 2024-03-14 | Th | No class - Spring Break |  |  |
| 2024-03-18 | M | Classifiers | MT-01 | HW-07 |
| 2024-03-20 | W | Assessing model performance |  |  |
| 2024-03-21 | Th | Journal Club |  |  |
| 2024-03-25 | M | Linear models I - Introduction | HW-07 | HW-08 |
| 2024-03-27 | W | Linear models II - Hypothesis testing |  |  |
| 2024-03-28 | Th | No class - Instructor at conference |  |  |
| 2024-04-01 | M | Linear models III - Multiple regression | HW-08 | HW-09 |
| 2024-04-02 | W | Linear models IV - Variable selection, regularized regression |  |
| 2024-04-03 | Th | Journal Club |  |  |
| 2024-04-08 | M | Logistic regression I | HW-09 | HW-10 |
| 2024-04-10 | W | Logistic regression II |  |  |
| 2024-04-11 | Th | Journal Club |  |  |
| 2024-04-15 | M | Linear mixed models – fixed / random effects | HW-10 | MT-02 |
| 2024-04-17 | W | Deviance |  |  |
| 2024-04-18 | Th | Journal Club: Ethics |  |  |
| 2024-04-22 | M | R packages |  | Final Package |
| 2024-04-24 | W | Dimensionality reduction |  |  |
| 2024-04-25 | Th | Journal Club: Stochastic Parrots |  |  |
| 2024-04-29 | M | Wrap-up | MT-02 |  |
| 2024-05-01 | W | No class - Finals |  |  |
| 2024-05-02 | Th | No class - Finals |  |  |
| 2024-05-06 | M | No class - Finals | Final Package |  |