PUBHLTH 490Z: Statistical Modeling in Health Data Science (3 credits)

Fall 2022 :: T/Th 11:30-12:45pm EST :: Hasbrouck Laboratory room 138

Course Head: <PROF_FULL_NAME>, Professor of Biostatistics, UMass - Amherst. <PROF_EMAIL>

TA: <TA_FULL_NAME>
Email: <TA_EMAIL>

TA: <TA_FULL_NAME>
Email: <TA_EMAIL>

Lectures: TuTh, 11:30am - 12:45pm EST

Course Website: <LINK_MOODLE>

Course Content

The aim of this course is to provide students with the modeling skills necessary to analyze and interpret relationships between variables in real world data. Specifically, they will develop the statistical and programming expertise necessary to analyze datasets with complex relationships between variables. Students will gain hands-on experience analyzing data using simple/multiple linear regression, logistic regression, Poisson regression and an introduction to machine learning. Students will learn how to build statistical models that can be used to describe and evaluate multidimensional relationships that exist in the real world. Students will work with the R statistical computing language and by the end of the course will require substantial independent programming. The course will not provide explicit or detailed training in R programming. To the extent possible, the course will draw on real datasets from biological and biomedical applications. This course is designed for students who are looking for a second course in applied statistics/biostatistics (e.g. beyond PUBHLTH 460), or an accelerated introduction to statistics and modern statistical computing.

Learning Goals *(By the end of the course students will be able to...)*

- use data to identify and distinguish patterns of randomness vs. non-randomness,
- understand and critique statistical model equations as representations of a given real-world setting,
- formulate, fit, and interpret statistical models to designed to answer specific scientific questions,
- weigh evidence for/against hypotheses about associations between variables,

– diagnose the appropriateness or “goodness-of-fit” of a given model,

– write concise, professional, and reproducible statistical analysis reports.

Lecture Notes
Lecture slides will be available on the course website.

Text
PUBHLTH 490Z will use the draft text *Introductory Statistics for the Life and Biomedical Sciences*, a revision of *OpenIntro Statistics, 3rd ed*. The revision is tailored for undergraduate students in the life sciences. Also included is an R companion for introductory statistics. PDF versions of both books will be freely available on the course website.

Reading
Brief reading assignments from the text or from notes will be assigned in advance of each lecture.

Computing
All computing for the course will be done using *R Studio*, a convenient interface to the statistical language R. Both *R Studio* and R are freely available and can be download to computers running either Mac OSX or Microsoft Windows.

Prerequisites
One of any of the following introductory stats courses taught at UMass: BIOSTAT 223, STAT 111, STAT 240, STAT 501, ResEcon 212, PSYCH 240, PUBHLTH 460. If you have not taken an intro stats course at UMass but still want to enroll in this course, you are encouraged to petition the instructor for permission, especially if any of the following apply: (a) you have taken AP Stats in high school, (b) you have taken a college-level intro stats course just not one of the ones listed above, or (c) you are confident in your quantitative skills and your ability to succeed in a fast-paced, advanced introductory course.
Additionally, prior programming experience with R or concurrent enrollment in BIOSTATS 597D(Introduction to Statistical Computing with R; Instructor: <NAME>) is required.

Expectations

This course will require you to work thoughtfully, carefully, and independently and will require substantial work outside of class time. Because we will be using a more project-driven approach in this course, with assignments that will build upon one another into a final product, it is vital that you do not fall behind. If you feel as though you are falling behind or starting to lose a handle on the content, I expect you to come talk to me either after class or during office hours so that I can help as much as I can to set you back on track. Please do not wait to talk to me if you start to fall behind.
I also expect you to devote substantial outside-of-class time to your work for this course, typically involving 5-10 hours per week. I anticipate that this work will be divided among:

- finishing in-class activities
- reading assigned articles and chapters
- reviewing your notes
- working on assignments
- conducting project work

- preparing for exams

Things you should expect from me:

- timely feedback on assignments and quizzes
- response to questions via email in < 2 working days (often sooner)
- attention to your questions related to coursework during office hours
- instruction in how to write, research, and debug R code

Things you should not expect from me:

- time for frequent non-office hour drop-in questions
- comments on a research project that is unrelated to your coursework
- writing your code for you or *extensive* debugging of your code

Course Grading

The course letter grade will be based on weekly problems sets, one midterm, and a final exam. The final course grades will use the standard boundaries of A-, A: 90 - 100; B-, B, B+: 80 - 90, etc.

Homework (40%): There will be approximately 5 homework assignments that you will complete over the course of the semester. Each assignment will have components that you will hand in for grading. Assignments and due dates will be posted in advance on the course website. The assignments will be graded. Most assignments will require you to submit a digital file with reproducible solutions, i.e. a Rmd file that reproduces your answers. Late homeworks will not be accepted under any circumstances. If a homework is not handed in on time, it will receive a grade of zero.

In-class quizzes (10%): There will be a in-class quizzes in this course that will cover material in Units 1 - 4 and will be announced periodically in the class.

Midterm exam (20%): There will be a in-class mid-term exam in this course that will cover material in Units 1 - 4 (see below).

Final exam (15%): There will be a take-home final exam in this course.

Project (15%): In the second half of the course, you will develop and write your own data story. A formal report on this project will be due at the end of the semester. Details will be provided in a separate hand-out.

Component	% of grade	When
Problem sets	40%	every 2 weeks
In-class quizzes (approx. 2)	10%	announced during course of semester
Midterm exam (in-class)	20%	Tuesday November 15, 2022
Course Project/Take home final	30%	Due Friday Dec. 16, 2022
Total	100%	

The teaching team will host office hours while students are working on the project and take-home exams to answer questions and provide support. Students will not be allowed to collaborate with each other in any way, and doing so will be treated as a violation of the Honor Code.

Course Schedule

This is a tentative course schedule and is subject to change with little or no notice.

Unit 0 RStudio and Rmarkdown
Unit 1 Random Variables
Normal distribution
Binomial distribution
Poisson distribution
Unit 2 Linear regression
Simple linear regression
Multiple linear regression
Unit 3 Logistic regression
Unit 4 Poisson regression
Unit 5 Introduction to machine learning
Final take home exam + course project

Council on Education for Public Health (CEPH) Course Competencies

- Distinguish among the different measurement scales and the implications for selection of statistical methods to be used based on these distinctions.
- Describe conceptual frameworks (statistical literacy) in biostatistics
- Apply biostatistical methods to the design of studies in public health.
- Use computers to appropriately store, manage, manipulate and process data for a research study using modern software.
- Apply descriptive techniques commonly used to summarize public health data.
- Describe the basic concepts of probability, random variation and selected, commonly used, probability distributions.
- Select and perform the appropriate descriptive and inferential statistical methods in selected basic study design settings.
- Describe appropriate methodological alternatives to commonly used statistical methods when assumptions are violated.
- Integrate analysis strategies in biostatistics with principles and issues in epidemiology. literature
- Develop written and oral presentations based on statistical analyses for both public health professionals and educated lay audiences.
- Apply statistical methods to solve problems in the health sciences and carry out theoretical research in statistical methodology.

Academic Integrity:

The UMass-Amherst Academic Honesty statement from the Academic Regulations 2006-2007, pp. 7-8 states:

> Intellectual honesty requires that students demonstrate their own learning during examinations and other academic exercises, and that other sources of information or knowledge be appropriately credited. Scholarship depends upon the reliability of information and reference in the work of others. Student work at the University may be analyzed for originality of content. Such analysis may be done electronically or by other means. Student work may also be included in a database for the purpose of checking for possible plagiarized content in future student submissions. No form of cheating, plagiarism, fabrication, or facilitating dishonesty will be condoned in the University community.

More information about the Honor Code as well as resources for students may be found at http://www.umass.edu/dean_students/codeofconduct.

Disability Statement

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus. If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website.

Student Resources

The following are our regular office hours:

- <PROF_FULL_NAME> (Thursdays: 10:15am-11:15am EST on Zoom or by appointment)
- <TA_FULL_NAME> (Wednesdays: 11:00am - 12:00pm on Zoom or by appointment)
- <TA_FULL_NAME> (Mondays: 1:00pm - 2:00pm on Zoom or by appointment)

Please feel free to stop by with your questions or your thoughts about the course. Additionally, you are welcome to send questions via email to any of the teaching staff.

All of the commands you will need in R will be demonstrated in lecture or section, but you may find the following summaries of R commands helpful.

- *The R Cookbook (TRC)*, by P. Teetor.
- *Getting Started with R Studio (GSRStudio)*, by J. Verzani.

Acknowledgements We are grateful to Professor <NAME> and Ms. <NAME> (Harvard University) for course materials.