STAT 530, Applied Multivariate Statistics and Data Mining – Fall 2024

Instructor:
David Hitchcock, associate professor of statistics
202 LeConte College
Phone: 803-777-5346
Email: hitchcock@stat.sc.edu
Course Web Page: http://people.stat.sc.edu/hitchcock/stat530.html
(Also accessible via Blackboard)

Classes: Meeting Times: Tues-Thurs 10:05-11:20 a.m., LeConte College Room 224

Office Hours: Monday, Tuesday, Wednesday, Thursday 9:00-9:50 a.m. or by appointment

Textbooks:
An Introduction to Multivariate Analysis with R (2011), by Brian Everitt and Tolsten Holthorn, available as a free (possibly only via USC computers) download at
https://link.springer.com/book/10.1007/978-1-4419-9650-3
An Introduction to Statistical Learning with Applications in R (2013), by James, Witten, Hastie, and Tibshirani, available as free download at
https://www.statlearning.com/

Prerequisite: A grade of C or higher in STAT 515 or equivalent. *For this purpose*, courses equivalent to STAT 515 include PSYC 228 or 709; EDRM 710; STAT 205, 509, 512, 700, or 704; MGSC 291, 391 or 692; BIOS 700; ECON 436. See me regarding other questions about prerequisites.

Course Outline: Much of Chapters 1 – 6 of the Everitt textbook, and Chapters 4, 5, 8, 9 of the James et al. textbook. Topics covered include: summary statistics for multivariate data, multivariate data visualization, principal components analysis, exploratory factor analysis, multidimensional scaling and correspondence analysis, cluster analysis, classification and supervised learning, tree-based methods, support vector machines, cross-validation, and (time-permitting) MANOVA, multivariate regression.

Learning Outcomes:
Upon successful completion of this course, students should be able to:
- Determine which multivariate methods are appropriate for a given situation
- Understand the basic logic behind each method's construction
- Verify whether the assumptions needed to implement the methods are satisfied
- Analyze a data set using the methods in an appropriate software package
- Interpret the output for each of the methods

Class Lectures / Attendance Requirement:
You are urged to attend the lectures live on Tuesdays and Thursdays at 10:05-11:20 a.m., in LeConte Room 224. If you are forced to miss a class or if you would like to review material from a class, I plan to ATTEMPT TO record the lectures on Zoom and post them in Panopto in the Blackboard STAT 530 course page. This is not a replacement for attendance and some aspects of the class (for example, if I write occasional notes on the board/document camera) may not show up on the recorded version. In addition, the recording technology is new this year (no longer Blackboard Collaborate), so I make **no guarantees** that the recordings will always be successful. In short: Please come to class!

Since this is an in-person class, you are expected to attend at least 80% of the class sessions in person. Attendance will be taken each class, and your grade on the attendance component will be 1.25 times the percentage of class sessions that you attend live (with a maximum of 100% for the attendance grade). For example, if you attend 60% of the class sessions in person, your attendance grade (which is 5% of the overall course grade) will be 75%. If you attend 80% or more of the class sessions in person, your attendance grade will be 100%.

Homework: Homework problems will be assigned periodically. Each student's homework must be done independently. You may ask each other informal questions about the homework, but everyone is to do his/her own work. If homework is found to be copied, all students involved will receive a 0. Of course, you may always ask me questions about the homework. [To be clearer, students can ask each other informal ORAL questions about homework, but **cannot look at or copy each other's homework papers**. All submitted homework must be their own work.] Note that for the take-home part of the midterm, you MAY NOT talk to each other about the problems at all.

Graduate Students: The university requires that 500-level classes be more rigorous for graduate students than for undergraduates. Homeworks will occasionally involve extra problems that will be required only for graduate students. These will be graded separately for graduate students, and a grade of 70% or higher must be earned on the combined extra sections across the problem sets or a letter grade (A→ B, B+→ C+, etc…) penalty will be applied to the final course grade.

Disabilities: Any student with a documented disability should contact the Student Disability Resource Center at 777-6142 to make arrangements for appropriate accommodations.

Exams: There will be one midterm exam with an in-class part (on Thursday, Oct. 10, 2024) and a take-home part which will be due a few days after that. There will be an in-person final exam on Tuesday, December 10 from 9:00-11:30 a.m. **You are not allowed to receive help from *anyone except me* on the take-home portion of the midterm exam.** For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.

Data Analysis Project: The project will be due near the end of the semester and will involve collecting or obtaining a real data set and analyzing it using the methods discussed in this class. There will be the option of working in teams or individually. More information will be given out later in class.

Grading: The course grade will be based on attendance (5%), homework average (20%), the midterm exam (30%), the project (15%), and the final exam (30%). The overall course average will result in the following grades: 90-100 = A, 87-89 = B+, 80-86 = B, 77-79 = C+, 70-76 = C, 67-69 = D+, 60-66 = D, 59 and below = F.

Computing: Use of a computer is required for the analysis of multivariate data. The examples in class will be done using R. No previous knowledge of R is assumed. Everyone is encouraged to download a free copy of R (see the course page for downloading instructions). All necessary analyses will be able to be done in R, at least --- you may use other software if it accomplishes the appropriate task and if you know how to use it.