Modernizing an Undergraduate Multivariate Statistics Class

David B. Hitchcock – University of South Carolina
Xiaoyan Lin – University of South Carolina
Brian Habing – University of South Carolina

August 4, 2016
STAT 530 at the University of South Carolina: Traditionally has been an applied multivariate analysis class.

For the last decade or so, it has been taught as a distributed learning class (online availability)

At South Carolina, 500-level courses are open to both undergraduates and grad students (must be “more rigorous” for grad students).
Audience of the Course

- A mix of statistics undergraduate majors, statistics M.S. and M.A.S. students, and students from other departments.

- Upcoming Fall: 25 UNDERGRADS: 14 stat majors, 3 computer science, 3 experimental psychology*, 1 each: sports and entertainment management*, math, biology, non-degree [* = a couple have stat minors]

- Upcoming Fall: 15 GRAD STUDENTS: 5 stat M.S. students, 6 stat M.A.S. students, 1 economics M.A., 1 sports and entertainment management Ph.D., 1 anthropology Ph.D., 1 chem. engineering Ph.D.

- Past years: 2014: 11 undergrads (6 stat majors), 17 grad students (4 M.A.S., the rest outside stat including 4 Business)
Online component

- Lectures recorded in a “distance studio”.
- Students have the option of (1) attending live, (2) watching live online, (3) watching recording at a later time.
- Most do not attend live.
- Students should be able to complete the course even if they can never come to campus.
Why Change the Old Course?

Data Science concept:

- Joint mini-program with Computer Science
- Will consist of 4 courses emphasizing:
  1. Data Visualization
  2. Advanced Statistical Methods/Models
  3. Supervised Learning (and Related Topics)
  4. Big Data Processing

- The CS department will handle (1) and (4), but a modernized Multivariate analysis course could address (3).

- Pick topics that will better suit students’ career needs
What to Add

- Modern Classification Methods: Support Vector Machine, Classification Trees, Random Forests, Logistic Regression, K-Nearest Neighbors
- Tree-Based Methods: Classification and Regression Trees, Random Forests, Bagging and Boosting
- Cross-Validation
What to Remove

- Some older classical (but less-used nowadays) methods
- Canonical correlation analysis
- Linear discriminant analysis (except in comparison to modern classification methods)
- Multivariate Regression? MANOVA?
- Will depend on how much time we have.
Question: How many classification methods do students need to know?

- Should students learn several “modern” classification methods?
- Are they better off learning one or two (leaving time to learn methods to solve other problems)?
- Classification is a “common” data analysis problem, while some other methods might be less commonly encountered.
What Students will be able to do when done

- Understand assumptions, limitations, and implementations of methods that we have covered
- Be comfortable using R for analysis of fairly large data sets
- Know how to pick right method to answer research question of interest
- Know how to compare competing methods in terms of, e.g., prediction success
- Main thing: Be excited and confident about using modern tools to explore data, predict, understand patterns, etc.
Different Audiences: Undergraduate vs. Graduate Students

- Technically this is an undergraduate course with minimal prerequisites, so we have to be careful with how much advanced material we give students.
- Graduate students make up a sizable part of the class too.
- They may want to learn more methods that they can use immediately in their research or work.
- A compromise: Give an overview and examples of numerous methods, but limit the amount of depth/background.
- Or: Present the background theory for those who want that, but don’t test students on it?
In the past, have used Everitt's (2005) *An R and S-PLUS Companion to Multivariate Analysis*

Everitt and Hothorn (2011) have a newer book that updates the topics in Everitt (2005): removes MANOVA, linear discriminant analysis.

For this fall, we are planning to use parts of the Everitt (2005) text along with the freely downloadable *An Introduction to Statistical Learning with Applications in R* by James, Witten, Hastie, and Tibshirani (2013).

The newer topics more related to data mining are covered in JWHT (2013).

In future semesters, may adjust textbook choices.
Structure of Assignments/Exams

- All assignments are take-home (distance format lends itself to this)
- Several smaller homework assignments to practice new methods
- Take-home midterm exam (mini-project)
- Take-home final exam (mini-project)
- For the exams, students are given somewhat large data sets with a series of somewhat open-ended research questions.
- Students should decide what analyses, graphics, etc., are called for and implement them.
- Students must hand in a formal, typed report.
Conclusions

- Balance of modern and classical methods is needed, but I think classical ("old-fashioned") methods that are rarely used should be deemphasized or phased out.
- Giving students practice in doing data analysis themselves and writing reports is very critical.
- For this level of class, the emphasis is on application, practical modeling choices, and computation.
- Just enough theory is given to provide intuitive justification for the approaches.
- The modernization/restructuring of the course is still ongoing.
- We may make different adjustments based on how the class goes this fall — or even based on comments I get from you!
References

- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.