STAT 460 – Applied Multivariate Analysis – Spring 2015

Instructor

	Office	Email	Office Hours
Darren Homrighausen	Stat 204	darrenho@stat.colostate.edu	Tuesday 11:00 or by Appointment

Lecture	MWF 10:00 AM – 10:50 AM Engineering B 101
Required Text	*Class notes*
Web Site	www.stat.colostate.edu/∼darrenho/AMA
Prerequisites	STAT340 and an undergraduate linear algebra class
Expectations	As this is a 3 credit class, there is a CSU expectation that you spend 6 hours outside class on homeworks and review each week.
Teaching Philosophy	I feel like students too often address school from the viewpoint of what you "have" to do or what the professor is "making" you do. This class, and everything that happens within it, is all for you. Realistically, I'm the one who "has" to do things; you "get" to do them. – Make it worth your time and monetary investment.

Most (essentially all) data is multivariate. *Applied Multivariate Analysis* is a topic meant to provide tools for visualizing, exploring and analyzing this type of data. Applications of multivariate statistics are happening all around you - and if they are done well, they may sometimes even go unnoticed. How does Google web search work? How does Shazam recognize a song playing in the background? How does Netflix recommend movies to each of its users? How could we predict whether or not a person will develop breast cancer based on genetic information? Can we automatically interpret incoming email to label it as spam? An expert's answer to any one of these questions may very well contain enough material to fill its own course, but basic answers stem from the principles of multivariate analysis.

Multivariate statistics involves a good deal of both applied work (programming, problem solving, data analysis) and theoretical work (learning, understanding, and evaluating methodologies). We will try to make the class as applied as possible. However, there are necessary detours into the more technical aspects.

Upon completing this course, you should be able to tackle modern multivariate problems by: (1) selecting the appropriate methods and justifying your choices; (2) implementing these methods pro-

grammatically (using, say, the R programming language) and evaluating your results; (3) explaining your results to a researcher outside of statistics.

Administrative Remarks

Honor Pledge

This class operates under the tenets of the CSU honor pledge: *I will not give, receive, or use any unauthorized assistance.*

Lectures

Attendance. Attendance is mandatory. You can do it.

Software

R. In this class you will be provided the opportunity to work with `R`; a widely used statistical computing platform. It is free, moderately well documented, and has a vast user community (download it from `www.r-project.org`).

Homework and Tests

Homework. There will be approximately 13 weekly homework assignments. *Homework assignments that are turned in late will be assessed a 30% penalty, regardless of the reason they are late! No homeworks will be accepted more than 24 hours after the due date/time.* Feel free to discuss homework assignments with others, but realize that the work you hand in must be your own. **VERY, VERY IMPORTANT:** For your submitted homeworks, you will be submitting code, output, and plots. Your code must be neat and readable. Only include output that is directly related to the question you are answering. Plots must be labelled (x-axis, y-axis, and title). Any deviations from this protocal may result in a deduction in points.

Tests. There will be an in class midterm April 3^{rd}. The specifics of the test, such as topics and permitted materials, will be discussed at a later time.

Final Policy. Instead of a final, we will have a final project which will include you finding a data set of interest to you and analyzing it using the methods discussed in this class. This project will have three parts: a write-up of your data set and your results, a class presentation, and a document summarizing everyone else's presentation. The details of the project will be fleshed out as the semester progresses. **VERY, VERY IMPORTANT:** Your project grade will be severely and adversely affected by not attending and summarizing other people's presentations. I will be keeping track.

Grade Rubric. Homework = 60%, Midterm = 15%, and Final = 25%. At a minimum, you'll receive a grade as per the usual letter grade partitioning of [0%, 100%].

Miscellaneous

Disability Resources. If you require a special accommodation, such as needing more time to finish exams, contact me **and** disability services.

Email. When sending email, please put "STAT460" at the beginning of the subject line so that I know the message is not spam.

Class Schedule

	Topics
0	Linear Algebra and Probability Review: matrix algebra, Singular Value Decomposition, Expectation and Distributions of vector-valued random variables, Bias and Variance of Estimators.
1	Regression Methods: Linear Regression, model selection, regularization (Ridge regression, Lasso, LARS).
2	Classification: Discriminant Analysis, Support Vector Machines, tree-based methods
3	Derived Inputs: Principal Components, Partial Least Squares, Diffusion Maps/Nonlinear Embeddings/Manifold learning, kernelization
4	Clustering: Proximity Matrices, K-means, Hierarchical clustering
5	Additional Topics, including: Factor Analysis, Text Analysis, ...

Note: This schedule is a general guide. We can (and probably will) deviate from this at some point. Additionally, if there is some relevant topic you'd like to cover, let me know.