Applied Data Analysis

TTP 289A-006 (CRN 90577)

Course Details:

Quarter:	Spring 2019
When:	Tues/Thurs: 1:10-3:00pm
Instructor:	Alan Jenn (ajenn@ucdavis.edu)
Eligibility:	Graduate level
Classroom:	1120 Hart
Number of Units:	4
Grading:	Letter graded

Course Description:

This course aims to provide students with the resources needed to examine, parse, and analyze datasets (with a specific aim for answering research questions). This data analysis course covers a variety of concepts across disciplines of economics, statistics, and machine learning but with a specific emphasis on application. All techniques will be taught through practical examples of real-world datasets with enough breadth to understand the most critical concepts behind various analysis techniques.

The concepts in the class include the exploration of data, gathering and cleaning of data. The course delves into basic data analysis operations, including basics of examining and inspecting data (identifying data types, dealing with missing data and outliers, maintaining data integrity). We will cover a range of regression analysis including parametric (OLS), semi-parametric (logistic), and non-parametric (GLM, kernel regressions) regressions. Lastly, we will apply the learned techniques to real data. We will cover a variety of datasets as examples (see potential datasets below) to demonstrate how to use the software tools.

Prerequisites:

None, programming and statistics/linear algebra background would be helpful but not required.

Course requirements:

1. Course participation: Students are required to actively participate in class discussion, this means attending class and interacting. Due to the hands-on nature of the course, participation will be the largest determinant of the final grade at 30% of the total.
2. Homework: Students will gain practical experience by analyzing real world data and applying learned techniques from lectures. The homework will replicate real world research: less as structured assignments and rather as open-ended practical problems that students may encounter as researchers. The homework will be worth 40% of the total.
3. Final project: Students will be required to select a dataset of their choice and analyze it using techniques learned throughout the class with a particular emphasis on maintaining an overarching research idea (as opposed to a series of disparate analyses). Projects will be conducted individually and are worth 30% of the final grade.

Grading:

Class participation	30%
Homework	40%
Final project	30%

Assignments:

There will be four different homework assignments, spaced throughout the quarter. The assignments will be relatively open-ended and aimed at replicating issues that students will encounter as they conduct research in graduate school. The assignments will not only help to apply techniques learned in class but to get students to think critically about how to approach data analysis. Late homework will be accepted with points deducted (25% deduction within 1 week, 50% deduction the next week).

Project:

The final project will be similar to the assignments but larger in scope. Students will be expected to conduct a cohesive and deep analysis of a dataset of their choosing using techniques learned throughout the course. Grading will be based on the accuracy and breadth of the analysis, as well as how the analysis would stand to critique in the academic realm.

Plagiarism:

"Plagiarism" means using the words or ideas of another without giving appropriate credit. Even if the student paraphrases the ideas in his/her own words, the source must be cited. If exact words are used, the student must put the words in quotation marks and cite the source. Students are responsible for knowing what plagiarism is and avoiding it. Be particularly careful about copying and pasting information from the Internet - materials used from Internet sources must be quoted and cited just like information from other sources. Students must also be aware that copying or adapting pictures, charts, computer programs or code, music, or data without citing sources and indicating that the material has been copied or adapted is plagiarism. It may also be copyright infringement.[1]

Course background:

The motivation behind this course is to provide students a practical applied course that teaches fundamental skills necessary for research (both in design and approach) in graduate school. The course draws on my background as a researcher and condenses an extensive set of classes from my graduate studies:

- Applied Data Analysis *(Department of Engineering and Public Policy)*
- Econometrics I *(Economics Department)*
- Econometrics II *(Economics Department)*
- Machine Learning *(School of Computer Sciences)*
- Decision Tools for Engineering Design and Entrepreneurship *(Mechanical Engineering Department)*
- Advanced Data Analysis *(Statistics Department)*
- Linear Regression *(Statistics Department)*
- Fundamentals of Programming *(School of Computer Sciences)*

[1] http://sja.ucdavis.edu/faq.html#20

Week	Date	Lecture	HW	Title	Topic
1	4/2/2019	1		Introduction to data analysis	Course overview, overview of data analysis, tools of the trade
	4/4/2019	2	HW 1 assigned	Fundamentals of programming	Crash course in programming (R), part 1: comparing to Excel, cleaning data, working with data, logic tables, functions, big O
2	4/9/2019	3		Approaching data	Crash course in programming (R), part 2: for loops, data types, data frames
	4/11/2019	4		Introduction to linear regression	Basics of regression: why least squares?, interpretation; model specification, standard error, coefficient of determination
3	4/16/2019	5	HW 1 due	Theory and practice of linear regression	Model comparison, F-test, fitted values, residuals, mean squared error
	4/18/2019	6	HW 2 assigned	Linear regression diagnostics	Outliers: leverage, influence; Multivariate regression, log-transformation
4	4/23/2019	7		Linear regression prediction	Dummy variables, interaction variables, cross-validation, confidence intervals
	4/25/2019	8		Linear regression causality	Regression as a tool in economics vs statistics/machine learning; correlation vs causation, directed acyclic graphs
5	4/30/2019	9	HW 2 due	Linear regression causality	Causal relationships: instrumental variables
	5/2/2019	10	HW 3 assigned	Non-parametric approaches	Importance of functional form: nonlinear regressions application, precision vs accuracy
6	5/7/2019	11		Non-parametric approaches	Nonlinear regressions: *k*-nearest neighbors, kernel regression, splines, kernels, general additive models
	5/9/2019	12		Semi-parametric approaches	Logistic function, odds-ratio, irrelevance of independent alternatives, choice modeling
7	5/14/2019	13	HW 3 due	Semi-parametric approaches	Multinomial logistic regression, model interpretation, probability outcome plots
	5/16/2019	14	HW 4 assigned	Introduction to machine learning	Core-concepts; supervised vs unsupervised learning; understanding relationship between economics, statistics, and machine learning
8	5/21/2019	15		Break for project work	
	5/23/2019	16		Break for project work	
9	5/28/2019	17	HW 4 due	Unsupervised learning	Introduction to classification; clustering: k-means, latent class/profile analysis, decision trees, neural networks, other algorithms
	5/30/2019	18		Data analysis: practical advice	Common pitfalls and important considerations; bringing techniques and methods together in a cohesive analysis; thinking skeptically: critiquing work, p-hacking
10	6/4/2019	19		Project presentations	
	6/6/2019	20		Project presentations	