# Applied Data Analysis

## TTP 289A-006 (CRN 82624)

**Course Details:**

Quarter: Spring 2020

When: Tues/Thurs: 10am-12pm

Instructor: Alan Jenn (ajenn@ucdavis.edu)

Eligibility: Graduate level

Classroom: 3 Wellman

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]](#footnote-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)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week** | **Date** | **Lecture** | **HW** | **Title** | **Topic** |
| 1 | 3/31/2020 | 1 |  | Introduction to data analysis | Course overview, overview of data analysis, tools of the trade |
|  | 4/2/2020 | 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 | 4/7/2020 | 3 |  | Approaching data | Crash course in programming (R), part 2: for loops, data types, data frames |
|  | 4/9/2020 | 4 |  | Introduction to linear regression | Basics of regression: why least squares?, interpretation; model specification, standard error, coefficient of determination |
| 3 | 4/14/2020 | 5 | HW 1 due | Theory and practice of linear regression | Model comparison, F-test, fitted values, residuals, mean squared error |
|  | 4/16/2020 | 6 | HW 2 assigned | Linear regression diagnostics | Outliers: leverage, influence; Multivariate regression, log-transformation |
| 4 | 4/21/2020 | 7 |  | Linear regression prediction | Dummy variables, interaction variables, cross-validation, confidence intervals |
|  | 4/23/2020 | 8 |  | Model selection | Model selection criteria: AIC, BIC, Adj R2, Stepwise Selection |
| 5 | 4/28/2020 | 9 | HW 2 due | Introduction to non-linear regression | Bias-variance tradeoff, regularization: ridge and lasso regression |
|  | 4/30/2020 | 10 | HW 3 assigned | Causal relationships, part 1 | Directed acyclic graphs, identification strategies |
| 6 | 5/5/2020 | 11 |  | Causal relationships, part 2 and dimension reduction | Instrumental variables and 2SLS, principle component analysis |
|  | 5/7/2020 | 12 |  | Semi-parametric approaches; classification | Logistic function and regression, odds-ratio, irrelevance of independent alternatives, choice modeling, multinomial logistic regression, model interpretation, probability outcome plots |
| 7 | 5/12/2020 | 13 | HW 3 due | Other classification examples | Linear discriminant analysis, *k*-nearest neighbors |
|  | 5/14/2020 | 14 | HW 4 assigned | Non-parametric approaches | Splines, natural splines, smoothing splines, generalized additive models |
| 8 | 5/19/2020 | 15 |  | Non-parametric approaches continued | Kernel regression, decision trees |
|  | 5/21/2020 | 16 |  | Introduction to machine learning | Neural networks, deep learning |
| 9 | 5/26/2020 | 17 | HW 4 due | No class | Work on presentations |
|  | 5/28/2020 | 18 |  | Data analysis principles | Parametric vs semi-parametric vs non-parametric, prediction vs causality, classification; practical approaches to data analysis; interpretation |
| 10 | 6/2/2020 | 19 |  | No class | Work on presentations |
|  | 6/4/2020 | 20 |  | No class | Work on presentations |
| 11 | 6/9/2020 | 21 |  | Project presentations |  |
|  | 6/11/2020 | 22 |  | Project presentations |  |

1. http://sja.ucdavis.edu/faq.html#20 [↑](#footnote-ref-1)