ECON 296-003
Understanding Data
Spring 2017
Department of Economics
George Mason University

Professor: Thomas Stratmann
Office: Carow Hall
Phone: (703) 993-4920
Email: tstratma@gmu.edu
Class: Thursdays, 4:30 pm - 7:10 pm, Robinson Hall A205 Jan 23, 2017 - May 17, 2017
Office Hours: TBA

Course Description

ECON 296 is a hands-on course with the focus on using a novel and intuitive, yet rigorous approach to teach students the fundamentals of quantitative reasoning and data exploration. Methods of analysis include descriptive statistics, testing for substantially important and statistically significant associations among continuous and categorical variables, testing for differences of means and independence, and correlation analysis. Students learn how to use cloud-based statistical software to uncover and interpret systematic patterns in a variety of data sets drawn from the social sciences and humanities.

Course Objectives

- Students will be able to analyze and interpret data, and be able to draw inferences from the data.
- Students will be able to model relationships between variables, and make quantitative predictions using these models.
- Students will be able to communicate statistical information, including presenting graphical representations of data and interpreting regression estimates.
- Students will be able to address real world problems and questions through analysis of data.

Textbook

None

Homework/Midterm/Final

Lecture notes, data sets, and problem sets will available on Blackboard. Correct answers to problem sets and reading assignments will be posted on this site.

The midterm exam will be held on March 23. This exam is a closed book exam. There will be no makeup midterm. If you miss the midterm with a valid excuse, its weight will be shifted to the final.

The final will be cumulative. Please refer to https://registrar.gmu.edu/calendars/spring-2017/final-exam/ regarding the date for the final exam.

Please familiarize yourself with the Honor Code, http://www.gmu.edu/catalog/apolicies/. Suspected cases of academic dishonesty including plagiarism will be sent immediately to the Honor Committee.

I will not accept late homework assignments.

Applied Computing
This course will include applied computing using Excel and DataSplash. You will learn how to use these programs and solve statistical problems through learning-by-doing. In either case, your work must be your own. Thus, please don't hand in someone else's work product.

For assigned problem sets we will use a combination of Excel (primarily for bar charts and pie charts) and DataSplash. Excel may be accessed at some of the different computer labs across campus. DataSplash is a cloud based tool that may be accessed http://www.datasplash.com.

DataSplash offers an interactive, intuitive, and instant approach to data exploration and visualization. Data can simply be imported into the software and with just a few clicks you can view summary statistics, regression results, produce graphical representations of your data, and much more.

Course Outline

Section 1: Describing and Visualizing Data

- Introduction to basic summary statistics: mean, median and mode.
- Spread of data: Max/Min, interquartile range, and variance
- Distribution of data: normality, skewness, outliers.
- Graphical representation of data: Scatter plot, bar chart, pie chart, etc.
- Correlation between data
- Types of Data: Cross-Sectional, Time-series, Panel

Section 2: Hypothesis Testing

- Correlation vs. Causation
- Research/Experimental Design
    - Treatment vs control group
    - Sources of error
- Comparison of means : T-test
- Chi-Squared test of independence

Section 3: Simple Regression

- Introduce Bivariate regression
    - Interpretation of coefficients
    - Visual representation best-fit-line
- Goodness of Fit
- Residuals
- Prediction using estimated model
- Relationship between regression and t-test
    - Dummy Variables vs Continuous Variables
- P-value

Section 4: Multivariable Regression

- Omitted variable bias
- Interpreting the multi-variable model
- Multi-collinearity

Section 5: Big Data

- A/B testing
- Machine Learning

Grading

Problem Sets: 50%
Midterm: 20%
Final: 30%

After having taking this course you know some answers to these commonly asked questions.

- where do I find data sets? (which online sources are trustworthy, how to scrape data off the web, collecting your own data)
- what makes for a good data set to explore? (when is it too small, too large, when is it too selective, when is it too broad)
- how do I formulate a testable hypothesis? (qualitative versus quantitative assertions, i.e. sign versus magnitude of difference)
- in multiple regression, what does "all else equal" mean? what does "controlling for x" mean?
- how to deal with missing data? (missing at random, selection on x, selection on y, should you drop missing observations or control for them [i.e. add a dummy for "value is missing"] or impute missing values?)
- how to deal with measurement error?
- how to do a survey? (e.g. minimum sample size needed to detect an effect / number needed to treat, ask one or multiple questions about the same issue)
- which questions should you ask to tell if an intervention / program works?

- how to report numerical results, e.g. relative versus absolute change, how to convey uncertainty (margin of error / standard error / p-value) around estimates , percent versus percentage point difference, statistical significance versus economic significance, explained variation / R-squared
- in multiple regression, how should you select your regressors (focus on R-squared? worry about multicollinearity? use substantive knowledge?)
- if an effect is not statistically significant, what does it mean? (there's really no effect or your sample is too small -- difference between evidence of absence versus absence of evidence)
- how to talk about your findings (e.g. blog post, radio interview, talk show interview, advertising copy, medical journal abstract)