14.310x Data Analysis for Social Scientists Syllabus

Esther Duflo
Professor of Economics, MIT

Sara Fisher Ellison
Senior Lecturer of Economics, MIT

Course Website:

https://www.edx.org/course/data-analysis-social-scientists-mitx-14-310x

Course Description

This course introduces methods for harnessing data to answer questions of cultural, social, economic, and policy interest. We will start with essential notions of probability and statistics. We will proceed to cover techniques in modern data analysis: regression and econometrics, design of experiments, randomized control trials (and A/B testing), machine learning, data visualization. We will illustrate these concepts with applications drawn from real world examples and frontier research. Finally, we will provide instruction on the use of the statistical package R, and opportunities for students to perform self-directed empirical analyses. Students taking the graduate version will complete additional assignments. No prior preparation in probability and statistics is required, but familiarity with basic algebra and calculus is assumed.

Prerequisites

Math: You should be prepared to keep up with an approach to economics that is somewhat mathematical. We suggest that you have taken high school calculus or the equivalent. We will use algebra in the lectures, problem sets, and exams.

Lectures

The material for each topic will be posted weekly, and you should keep pace with the rest of the class. There will be about two lectures per week. You will have access to videos of the lecture presented in short segments (8-10 minutes on average), followed by finger exercises to test your understanding of the material. You will also have access to the presentation slides to follow along during the lecture.

Time Commitment

The minimum commitment will be approximately 8-12 hours per week for watching the lectures, doing the readings, and completing the assignments.

Assignments and Grading Scheme

- For each unit of the course, there will be a series of finger exercise questions after each video segment and a homework assignment. Homework assignments will be released on Mondays along with the videos, and will be due the following Sunday, giving you about two weeks to complete the assignment. Problem sets will contain empirical exercises, theory exercises, and short-answer written questions. We will provide resources to help you learn how to use the statistical software R to complete these exercises. Problem sets for this class are very important, and can be somewhat time-consuming.
- In addition, there will be a final exam which you will have about two and a half days to complete. Please see the Schedule and the Grading Policy document for further information.
- Students scoring at least 50% will earn a certificate. Grades are calculated as follows:
  - Homework Assignments: **45%**
  - Finger Exercises: **30%**
  - Final Exam: **25%**

Helpful Textbooks

There are no required texts for the course. We will draw on material from many sources. For the first half of the course, a book in probability and statistics could be useful for reference. Possible titles include Introduction to Mathematical Statistics and its Applications by Larsen and Marx, Probability and Statistics by DeGroot and Schervish or Statistical Theory by Lindgren. The first is probably the easiest and most discursive. The second is an excellent but somewhat more difficult book. The third is a great book for reference but doesn't offer much intuition. There is no text that will cover most of the second half of the course, but both Introductory Econometrics by Wooldridge and Introduction to Econometrics by Stock and Watson have some overlap with what we will do and could be useful references in the future.

Syllabus & Schedule

1. **MODULE 1: INTRODUCTION**

    - Introduction to the software R with exercises. Suggested resources for learning more on the web.
    - Introduction to the power of data and data analysis, overview of what will be covered in the course.

2. **MODULE 2: FUNDAMENTALS OF PROBABILITY, RANDOM VARIABLES, DISTRIBUTIONS AND JOINT DISTRIBUTIONS**

    - Basics of probability and introduction to random variables.
    - Discussion of distributions and joint distributions.

3. **MODULE 3: GATHERING AND COLLECTING DATA, ETHICS, AND KERNEL DENSITY ESTIMATES**

    - Introduction to collecting data through surveys, web scraping, and other data collection methods.
    - Principles and practical steps for protection of human subjects in research.
    - Discussion of kernel density estimates.

4. **MODULE 4: JOINT, MARGINAL, AND CONDITIONAL DISTRIBUTIONS & FUNCTIONS OF RANDOM VARIABLES**

    - Builds on the basics from module 2 to cover joint, marginal, and conditional distributions.
    - Similarly builds on the basics from module 2 to cover functions of random variables.

5. **MODULE 5: MOMENTS OF A RANDOM VARIABLE, APPLICATIONS TO AUCTIONS, & INTRO TO REGRESSION**

    - Discussion of moments of a distribution, expectation, and variance.
    - Application: application of some principles of probability to the analysis of auctions.
    - Basics of regression analysis.

6. **MODULE 6: SPECIAL DISTRIBUTIONS, THE SAMPLE MEAN, CENTRAL LIMIT THEOREM, AND ESTIMATION**

- Discussion of properties of special distribution with several examples.
- Statistics: Introduction to the sample mean, central limit theorem, and estimation.

7. **MODULE 7: ASSESSING AND DERIVING ESTIMATORS- CONFIDENCE INTERVALS AND HYPOTHESIS TESTING**

    - Deriving and assessing estimators.
    - Constructing and interpreting confidence intervals.
    - Introduction to hypothesis testing.

8. **MODULE 8: CAUSALITY, ANALYSING RANDOMIZED EXPERIMENTS, & NONPARAMETRIC REGRESSION**

    - Understanding randomization in the context of experimentation.
    - Introduction to nonparametric regression techniques.

9. **MODULE 9: SINGLE AND MULTIVARIATE LINEAR MODELS (2 Lectures)**

    - In-depth discussion of the linear model and the multivariate linear model

10. **MODULE 10: PRACTICAL ISSUES IN RUNNING REGRESSIONS, AND OMITTED VARIABLE BIAS**

    - Covariates, fixed effects, and other functional forms
    - Introduction to regression discontinuity design

11. **MODULE 11: INTRO TO MACHINE LEARNING AND DATA VISUALIZATION**

    - Introduction to the use of machine learning for prediction. Covers tuning and training
    - Principles of data visualization with examples of well-crafted visual presentations of data

12. **MODULE 12: ENDOGENEITY, INSTRUMENTAL VARIABLES, AND EXPERIMENTAL DESIGN**

    - Understanding the problem of endogeneity. Introduction to instrumental variables and two stage least squares, with a discussion of how to assess the validity of an instrument.

- Discussion of how to design and effective experiment, followed by an example from Indonesia.

**FINAL EXAM