14.310x: Data Analysis for Social Scientists

Intructors:

- Esther Duflo, Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics, Department of Economics (MIT)
- Sara Fisher Ellison Senior Lecturer, Economics (MIT)

Course Descriptions:

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

If you are interested in getting an overview of the content and exercises covered in this course, or eager t to find additional resources please check out our course preview. A score of 60% or above in the course previews indicates that you are ready to take the course, while a score below 60% indicates that you should further review the concepts covered before beginning the course.

Prerequisites:

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

This Course and the MicroMasters:

This course is now part of two independent MITx MicroMasters Programs. The MITx MicroMasters is a professional and academic credential for online learners from anywhere in the world. Learners who pass an integrated set of MITx graduate-level courses on edX, and one or more proctored exams, will earn a MicroMasters credential from MITx, and can then apply for an accelerated, on campus, graduate degree program at MIT or other top universities.

We want to ensure that if any of our learners are interested in joining either of these programs, they have a clear understanding of how each program works as the certificates for each program are not interchangeable. You can find more information below, and should you have any questions you can refer to our MicroMasters course page, the discussion forum posts, or reach out to us at micromasters-support@mit.edu.

First, for each of the MicroMasters programs you will need to pass this online course with a final grade of 50% or above. However, each program will then require different final assessments in order to earn a course certificate toward the full MicroMasters credential with details and logistics as follows:

The Data, Economics, and Development Policy (DEDP) Program:

If you are interested in pursuing the DEDP MicroMasters credential, you will need to create a MicroMasters Profile, pay your DEDP MicroMasters course fee, pass this online course (50% or above) and pass an additional in-person proctored exam. You will then have fulfilled one class towards the DEDP MicroMasters credential, and will receive a certificate. Program website: https://micromasters.mit.edu/dedp/.

The Statistics and Data Science (SDS) Program:

If you are interested in pursuing the DS MicroMasters credential, you will need to pass this online course (50% or above), enroll and verify for the assessment course 14.310Fx Data Analysis in Social Sciences-Assessment on edX by the specified deadline, and successfully pass the assessment course (60% or above). You will then have fulfilled this component of the SDS MicroMasters credential and will receive a certificate. Program website: https://micromasters.mit.edu/ds/.

Assignments and Grading Scheme:

For most weeks during the course, there will be a homework assignment that covers the main topics in that unit. Homework assignments will be released on Mondays along with the videos, and will be due Sunday. In addition, there will be a final exam. Please see the online calendar for further information. There is also a pdf schedule that you can download and keep for offline use.

Grades of the edX course are calculated as follows:

- Homework Assignments: **45%**
- Finger Exercises: **30%**
- Final Online Exam: **25%**

Students who are taking this course in pursuit of the **MicroMasters credential** will also have to sit an in-person, proctored exam.

To be eligible to register for the proctored exam, you first will have to pass the online component of the course on edX as outlined above. Your final MicroMasters course grade will be calculated as follows:

- edX Course: **40%**
- Proctored Exam: **60%**

Lectures and Time Commitment:

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. You will also have access to the lecture notes and presentation slides.

The minimum commitment will be approximately 12-14 hours per week for watching the lectures, doing the readings, and completing the assignments.

edX Honor Code Pledge:

By enrolling in a course on edX, you are joining a special worldwide community of learners. The aspiration of edX is to provide anyone with an internet connection access to courses from the best universities and institutions in the world and to provide our learners the best educational experience internet technology enables. You are a part of the community that will help edX achieve this goal. edX depends upon your motivation to learn the material and to do so with honesty and academic integrity. By enrolling in an edX course, you have agreed to the edX Honor Code, which means that you will:

- Complete all graded material (graded assignments and exams) with your own work and only your own work. You will not submit the work of any other person or have anyone else submit work under your name.
- Maintain only one user account and not let anyone else use your username and/or password. Having two user accounts registered in this course will constitute cheating.
- Not return to a previous semester of the course in order to utilize assignment or exam answers for the current semester.
- Not engage in any activity that would dishonestly improve your results, or improve or hurt the results of others.
- Not collaborate with anyone other than staff on the graded assignment or exam questions. This means comparing answers, working as teams, or sharing answers in any way. The Practice Problems and Quick Questions are designed for collaboration - not the Graded Assignments, midterms or final exams.
- Not post answers to problems that are being used to assess student performance.
- Always be polite and respectful when communicating across the platform (with other learners and the staff).

We will strictly enforce the honor code pledge. Students found violating this pledge will be dealt with directly. **If we become aware of any suspicious activity we reserve the right to remove credit, not award a certificate, revoke a certificate, ban from any and all DEDP courses as well as notify edX for other actions.** We take academic honesty very seriously at MIT. With the introduction of the MicroMasters Credential, the important of honesty in work has been elevated to a higher level than before and we will diligently monitor this.

Course Syllabus and Reading Assignments:

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.

Module One: 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.

Module Two: Fundamentals of Probability, Random Variables, Joint Distributions, and Collecting Data

- Basics of probability and introduction to random variables
- Discussion of distributions and joint distributions
- Introduction to collecting data through surveys, web scraping, and other data collection methods

Module Three: Describing Data, Joint and Conditional Distributions of Random Variables

- Principles and practical steps for protection of human subjects in research
- Discussion of kernel density estimates
- Builds on basics from module 2 to cover joint, marginal, and conditional distributions

Module Four: Joint, Marginal, and Conditional Distributions and Functions of Random Variables

- Similarly builds on the basics from module 2 to cover functions of random variables
- Discussion of moments of a distribution, expectation, and variance
- Basics of regression analysis
- Application: Application of some principles of probability to the analysis of auctions (optional)

Module Five: Special Distributions, The Sample Mean, The 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

Module Six: Assessing and Deriving Estimators, Confidence Intervals and Hypothesis Testing

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

Module Seven: Causality, Analyzing Randomized Experiments, and Nonparametric Regression

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

Module Eight: Single and Multivariate Linear Models

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

Module Nine: Practical Issues in Running Regressions and Omitted Variable Bias

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

Module Ten: Intro to Machine Learner 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

Module Eleven: 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 the effective experiment, followed by an example from Indonesia

***Week Eleven: Final Online Exam**