Introduction to Data Analysis and Econometrics

Yale University, Summer 2021

Updated on: November 25, 2020

ADMINISTRATIVE

INSTRUCTOR

Helen Pushkarskaya(helen.pushkarskaya@yale.edu)
Office: TBD Office Hour: TBD

TIME AND LOCATION

Lecture Time: TBD Lecture Location: TBD

TEACHING FELLOWS

TBD

TUTORS AND OTHER RESOURCES

TBD

WEBSITE

On Canvas at: TBD

ABOUT THE COURSE

This course will teach you how to judge quantitative information and how to use data to answer economic and social questions. We will cover four areas:

1. *Probability*, the study of uncertainty, such as the uncertainty faced byinvestors, insurers, and people in everyday life.
2. *Statistics*, the science of analyzing and interpreting data, such as what a marketing department might know about past consumer purchases.
3. *Linear regression*. A statistical method used to estimate the relationship between two or more variables.

4. *Causality*. When can statistical analysis make a claim about causation?

The prerequisites for this course are introductory microeconomics and familiarity with single variable calculus. This course fulfills the econometrics requirement for the economics major.

In most econometrics classes, mathematical methods are introduced and then, some time later, applied to a few examples. This class turns that around. We will focus on substantive questions from the start, and gradually introduce mathematical methods that will help us answer them.

By the end of the class, you will have gained several skills:

1. Be able to choose appropriate statistical methods to answer real-world questions.
2. Understand both the math and the intuition behind methods like linear regression and hypothesis testing.
3. Be able to apply these methods to analyze real data with a powerful statistical analysis package (R).

We will apply our skills to a range of topics in economics, including intergenerational mobility, energy and environmental economics, discrimination, development economics, public health, and finance.

GRADES

Introductory Economics Curve

The course follows the curve set for all introductory courses in the Economics Department, per the department guidelines.

Grade Components

I reserve the right to change this breakdown. Your grade will be based on the following components:

1. Participation (15%)

   You will be expected to attend the interactive lectures each week and participate in group discussions.

2. Online Quizzes (10%)

   You will have five short on-line quizzes, one each week. The quizzes will be primarily based on materials presented during pre-recorded lectures, they are meant as a quick review and generally quiz questions will be *less* difficult than exam questions.

   The quizzes are open book/open notes, but you cannot collaborate with other students on the quizzes or discuss the quizzes with other students until after their solutions are posted. The lowest quiz score will be dropped.

   Quizzes will be posted on Thursdays after class during the weeks where there is not an exam. Quizzes are due on the evening of the day after the assignment date.

3. Problem Sets (30%)

There will be 3 problem sets during the semester. These problem sets will be primarily empirical and based on research papers.

You may work in groups of up to four people on the problem sets, but you must turn in your own individual assignment. If you work in a group, you must indicate on your submission the other members of your group. Your problem set with the lowest grade will be dropped.

The problem sets must be submitted through the course web site by 12:00 pm (noon) on the due date (one week after the assignment date). Any problem set not submitted by noon on the due date will not be accepted without a note from your residential dean emailed to me before the due date.

We will grade a randomly selected subset of the problems from each assignment, check for completion of the other parts of the assignment, and post suggested solutions after the problem sets are due.

4. Midterm Exam (20%)

The midterm will be in-class on TBD. Exams are closed book, but you may bring one double-sided page of notes for the midterm.

5. Final Exam (25%)

The final exam is cumulative and is scheduled for TBD. Exams are closed book, but you may bring two double-sided pages of notes for the final.

Errors in Grading

If a student believes that there has been a mistake in their grading, the student must prepare a written statement describing in detail the mistake, which then should be emailed to us.

Changing assigned grades is extremely unlikely and reserved for clear errors made by ourselves. Re-grading will not be considered unless submitted in writing via email as described above.

LECTURES

The course includes two pre-recorded and two live/interactive lectures a week. Pre-recorded lectures will be posted online on Friday before the respective week on the course webpage. Students are expected to watch them by the specified date; they will be quizzed on materials from these pre-recorded lectures. Slides for live/ interactive lectures will be posted on the course webpage but are not designed as a substitute for attending lecture. If you cannot attend a particular lecture, please augment the lecture slides with the notes from a student who did attend the lecture.

If students choose to use laptops or tablets to take notes, they may not use them to access non-course-related websites, Apps, or email during class.

ACCEPTABLE USE POLICY

You are free to use any published materials (e.g., a textbook), in preparing Econ 117

assignments or for learning the material more generally. Similarly, you are free to use online resources such as stack overflow questions or R tutorials. You are also strongly encouraged to work with others in your class. This is particularly helpful for learning to program. Each person must turn in their own assignment.

The use of any solution materials prepared in a previous year for Econ 117 or Econ 131, other than materials distributed this academic year by the course faculty, is **strictly prohibited** and constitutes cheating. This includes 1) any notes, spreadsheets, or handouts distributed in a prior term of Econ 117 or Econ 131; and 2) any notes, solutions, or spreadsheets prepared by former students of Econ 117 or Econ 131, in either written or electronic form.

This policy means you should not solicit or use solutions to previous years' problem sets. The reason for this policy is that access to previous year's materials can create serious inequities between fellow students, and jeopardize the integrity of the academic environment. Any violation of this policy will be reported.

We do not tolerate cheating and plagiarism. Cheating or plagiarism will result in a 0 on the assignment and will be reported to the college dean. You are welcome to work together in groups up to 4, but you are required to submit your own write-up and your own code.

Please take precautions to avoid putting us in a situation where we are forced to decide if two documents are "too similar". As future researchers, consultants, bankers, entrepreneurs, etc, learning to do honest work in a timely manner is more important than getting everything correct.

If you are uncertain, please add proper citation. For example, if you relied heavily on a group-member's code for one part of an assignment, then you should make a footnote highlighting this fact. This may result in a slightly lower grade, but as long as proper credit is clearly given, it does not constitute cheating. The one exception to this rule is using past material from any previous version of this course.

SOFTWARE

Much of the course work in Econ 117 will involve analysis of data using R, an open source implementation of the object-oriented programming language S. It is widely used by applied statisticians and its libraries implement a wide variety of statistical and graphical techniques with applications to a range of disciplines, such as the agricultural and biological sciences, genetics, neuroscience and economics.

R can be downloaded from https://cran.r-project.org. We will provide some notes on the use of R as a part of the pre-recorded online materials. There are also many excellent and free R references available online, for example, *Econometrics in R* by G. Farnsworth that is available for free. If your time permits and you want to dig deeper, there are also more programming oriented references such as *An Introduction to R* by W. N. Venables, D. M. Smith and the R Core Team. However, I recommend learning by trial and error, as it is the most time efficient approach and sufficient for the type of coding problems that we will consider.

If you have never used R (and have never used another programming language), I recommend completing one or both of these online introduction tutorials:

- www.codeschool.com/courses/try-r (1-2 hours). This tutorial now requires a credit card to register for a free trial. If you cancel within 10 days, you will not be charged.
- www.datacamp.com (3-5 hours). This tutorial is free.

TEXTBOOKS

There is no required textbook for this course.

While there is no required textbook, you may consider some optional textbooks if you are having trouble following the material. An excellent econometrics textbook is Introduction to Econometrics, 2nd or 3rd edition, by Stock and Watson (Addison-Wesley, 2010). It's coverage of probability and statistics is somewhat rudimentary, but it's treatment of regression methods is excellent and the book should serve you well as a reference in the future. An alternative with less math but more intuition is Mastering Metrics by Angrist and Pischke (Princeton University Press, 2014).

For students without a strong mathematical background, you may also find the following (optional) text useful: Probability and Statistical Inference, 8th or 9th ed., by Robert Hogg, Elliot Tanis, and most recently Dale Zimmerman (Pearson, 2010 or 2015). Hogg et al provides much deeper coverage of the concepts covered in the first half of the course than does Stock and Watson. The most important method we will cover during the course is linear regression and I highly recommend Paul Allison's *Multiple Regression: A Primer*. The writing is extremely clear and he covers both the intuition and mathematics behind the method.

CLASSROOM POLICIES

- Students with disabilities should contact Judy York in the Resource Office on Disabilities, 203-432-2324 or talk to me during the beginning of the course to make sure all needs are accommodated.
- This class is committed to an inclusive learning environment. All students, teaching staff, and the professor are expected treat each other with respect and dignity at all time. This includes posts on Piazza.
- All community members should enjoy an environment free of any form of harassment, sexual misconduct, discrimination, or intimate partner violence. If you encounter sexual harassment, sexual misconduct, sexual assault, or discrimination based on race, color, religion, age, national origin, ancestry, sex, sexual orientation, gender identity, or disability please contact the Title IX Coordinator, Stephanie Spangler, at stephanie.spangler@yale.edu (203.432.4446) or any of the University Title IX Coordinators, who can be found at: http://provost.yale.edu/title-ix/coordinators

ACKNOWLEDGEMENTS

This class, including course structure, slides, and problem sets, was developed jointly with Edward Vytlacil, John Eric Humphries and Nicholas Ryan, building in large part from the course

Daniela Morar and Vitor Possebom taught at Yale in Summer 2019, which in turn was heavily influenced by the courses Douglas McKee taught at Yale in Fall 2015 and Professor Lanier Benkard taught at Yale in Fall 2010. We are extremely grateful to them for sharing their syllabus, lecture slides, assignments, handouts, exams, and advice.

CIP Code

Note that the economics program has changed its CIP code. The new CIP code (for Classification of Instructional Programs) by the National Center for Education Statistics at the Department of Education is 45.0603 (Econometrics and Quantitative Economics) rather than the old one 45.0601 (Economics, General).

SCHEDULE

The schedule is subject to change.

Week	Prerecorded materials	Interactive lecture	Materials	Problem sets
1		1. Introduction: Econometrics is Eating the World	Econometrics and data analysis underlying policy debates and changes in the economy. Motivating example of income inequality. Course skills: skepticism, R. Start with probability definitions.	Pset 1 Assigned
	1. Joint and Conditional Probability. Independence.		Joint probability and conditional probability; Bayes' rule; definition of independence.	
		2. Application of probability rules.	Applications in evictions data. Applications in mobility data. Applications to college and income mobility. **Quiz 1**	
	2. Random Variables (One at a Time)		Definition of random variable; discrete and continuous random variables; distributions and distribution functions; definitions of probability density function and cumulative distribution function. Examples from variables we will encounter: college attendance (discrete, dummy), firm formality, profits; pollution readings; natural gas prices. Using statistics to describe distributions: central tendency, dispersion/variance.	

2		3. Random Variables (Two At a Time)	Joint and marginal distributions. Conditional expectations and independence. Covariance. Rules for expectation and variance of linear combinations of random variables. Law of iterated expectations. Distribution of profits conditional on formal and not formal.	Pset 2 Assigned
	3. The Sample Is Not the Population		Arbitrary / empirical distributions vs. parametric families of distributions, e.g. binomial, e.g. exponential, e.g. Pareto with Piketty data. Idea of population and sample statistics. Estimators. Sample mean and unbiasedness.	
		4. The Central Limit Theorem	Idea of simulation. The sample mean. Illustrating the Central Limit Theorem for sample means. CLT in data visualization. Write random variable generating function and calculate sample means. **Quiz 2.**	
	4. Hypothesis Testing: Could This Have Happened by Chance?		Hypothesis tests based on sample means. Standard error, confidence intervals, and p-values. How these numbers appear in research; how they might appear in the news (margin of error); what you should infer if they do not appear.	
3		5. Selection and Causation	Rubin causal model and college attendance. Conditional means for discrete conditioning variables. RCTs with full compliance and partial compliance: the logic of selection into social programs and leading examples of RCTs on social policy. Wald estimator. Elusive nature of ATE with imperfect compliance.	
	5. Linear regression - Introduction		Need for simpler models: lack of data and sampling error, interpretation of parametric models. Ordinary least squares in univariate case. Formulae for slope and constant and reference to OLS objective function. Mincer regression example using OLS.	
	Midterm exam			
	6. Linear regression - Bivariate		R2 and goodness of fit. Running regressions in R and interpreting output. Specifying bivariate regressions. Functional form.	

4		6. Linear regression - Multivariate	Estimating multiple regressions. Omitted variables bias. OVB formula and interpretation. Choosing regression specifications, good vs. bad control variables. **Quiz 3.**	Pset 3 Assigned
	7. Regression modeling		Regression inference. Interpreting regression coefficients.	
		7. Fixed Effects and Difference-in- Differences	Application of multiple regression: difference- in-difference. Specification of difference-in- difference and fixed effects models. Identifying assumption and test for pre-trends. Leading cases: policy changes. Research examples. **Quiz 4.**	
	8. Regression Discontinuity		Application of multiple regression: regression discontinuity. Identifying assumption and tests for sorting. Best practices for running variable controls. Leading cases: administrative rules, test score cut-offs. Research examples covering a range of questions.	
5	9. Simultaneity, instrumental variables		Examples of simultaneity: demand and supply, education and earnings, etc. Discussion of error term in OLS and identifying assumption of exogeneity. Introduction to IV and 2SLS; two- step and one-step implementations of 2SLS.	
		8. Instrumental variables	Monte Carlo IV in a structural model of demand and supply. Compare OLS and IV results. Compare OLS and IV results in a real data set. **Quiz 5.**	
	10. Final Review		Final Review	
	Final Exam			