# Introduction to Data Analysis and Econometrics

**Yale University, Summer 2021**

Updated on: February 24, 2021

## ADMINISTRATIVE

### INSTRUCTOR

Vitor Possebom (vitoraugusto.possebom@yale.edu)
Office Hours: MWF 11:15am-12:15pm

### TIME AND LOCATION

Lecture Time: MWF 9:00am-11:15am
Lecture Location: Zoom

### TUTORS AND OTHER RESOURCES

The Yale Summer Session offers Economics and Writing tutoring (https://summer.yale.edu/admitted-students/summer-session-tutoring).

Economics Tutoring Time: W 5:00pm-6:00pm

Writing Tutoring Time: Th 6:30pm-8:00pm
Sat 1:30pm-3:00pm
M 11:30am-1:00pm

### WEBSITE

On Canvas at: TBD

On Piazza 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 five areas:

1. *Probability*: the study of uncertainty, such as the uncertainty faced by investors, 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, discrimination, development economics and finance.

## GRADES

Introductory Economics Curve

The course follows the department-mandated curve set for all introductory courses in the Economics Department.

Grade Components

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

1. Online Quizzes (10%)

You will have three short on-line quizzes, two before the midterm and one after. The quizzes 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 before their deadlines. Therefore, do *not* post on Piazza about the quiz before their deadline. The quizzes' solutions will not be posted, but there will be ungraded versions of all quizzes that can be taken as many times as you want. You should use them to review the material.

The lowest quiz score will be dropped.

Quizzes will be posted on Fridays after class and must be submitted through the course web site by 5:00pm on the next Monday. Any quiz not submitted by 5:00pm on the due date will not be accepted without a written explanation emailed to the instructor. The instructor may or may not accept a late quiz depending on the student's explanation.

- Quiz 1: posted 07/16, due 07/19
- Quiz 2: posted 07/23, due 07/26
- Quiz 3: posted 08/06, due 08/09

2. 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 will be assigned on Fridays after class and must be submitted through the course web site by 9:00am on the defined due date. Any problem set not submitted by 9:00am on the due date will not be accepted without a written explanation emailed to the instructor. The instructor may or may not accept a late problem set depending on the student's explanation.

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.

- posted 07/16, due 07/23
- posted 07/23, due 08/02
- posted 08/06, due 08/13

3. Midterm Exam (30%)

The midterm will be a 24-hour exam that will end at 5:00 pm on July 28th. Exams are open book, but you cannot ask for the help of any human being. You should be able to finish the exam in less than 3-hours, but you will have 24 hours to complete it so that we can accommodate people located in different time zones.

4. Final Exam (30%)

The final exam is cumulative and will be a 24-hour exam that will end at 5:00 pm on August 13th. Exams are open book, but you cannot ask for the help of any human being.

You should be able to finish the exam in less than 3-hours, but you will have 24 hours to complete it so that we can accommodate people located in different time zones.

Errors in Grading

If students believe that there has been a mistake in 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

Lectures will be interactive. Lecture slides 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.

A lecture's length is 2h15min. Some lectures will cover more than one topic and will be broken in 2 parts of 1h05min with a 5min-break between them. All lectures will be associated to an optional reading. Moreover, all lectures will be recorded and uploaded to our course's Canvas webpage.

## LABS

Labs are optional (but strongly recommended) assignments that apply the methods learned in class in R. We use the methods we learned in lecture to analyze real data and answer real research questions. The labs will be designed to help you with your problem sets because they will cover skills that you will need for the problem sets.

## PIAZZA

Piazza is a valuable resource that allows the students to communicate among themselves and with the instructor. Discussing a lecture topic or a question from the problem set with your peers is an important learning tool that is easily available through Piazza. Moreover, by asking questions to the instructor through Piazza, you can also help your classmates who may have the same question. For this reason, Piazza is the recommended way to contact the instructor, who can also be contacted via email.

## 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 stackoverflow questions or R tutorials. You are also strongly encouraged to work with others in your class. This is particularly helpful for learning to code. 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 department. 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 handouts on the use of `R` and the program documentation is excellent. There are also many excellent and free `R` references available online, for example, *Econometrics in R* by G. Farnsworth and and *Applied Econometrics with R* by Christian Kleiber and Achim Zeileis. 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), it might be helpful to complete one or both of these online introduction tutorials:

- https://www.pluralsight.com/courses/r-programming-fundamentals (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.

- https://www.datacamp.com/courses/free-introduction-to-r (3-5 hours). This first part of this tutorial is free and it is enough as a starting point.

## 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). Its coverage of probability and statistics is somewhat rudimentary, but its 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.

Moreover, I will recommend chapters from *Introduction to Econometrics with R* by Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer as optional reading for each lecture.

## CLASSROOM POLICIES

- Students with disabilities should contact Sarah Scott Chang 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.
- Attendance is mandatory. Even one absence can significantly affect the student's ability to keep up. However, the instructor can grant excuses if the student emails me with a reasonable explanation. If a student has multiple absences, it may be necessary to impose a cut restriction. Any further absences might result in the student's involuntary withdraw from the course.

## ACKNOWLEDGEMENTS

This class was developed jointly with Edward Vytlacil, John Eric Humphries, Nicholas Ryan and Daniela Morar, building in large part from the course Douglas McKee taught at Yale in Fall 2015, which in turn was heavily influenced by the course Professor Lanier Benkard taught at Yale in Fall 2010.

The course structure, slides, and problem sets are also influenced by discussions with Dr Rebecca Toseland and the course material from Professor Raj Chetty's Econ 45 at Stanford University and Professor Esther Duflo's 14.310 at MIT. Thank you also to Professor Lisa Kahn from Yale School of Management. We are extremely grateful to them for sharing their syllabus, lecture slides, assignments, handouts, exams, and advice. In addition, Majed Dodin and Eduardo Fraga prepared the data sets for this course and helped tremendously with the construction of the problem sets and revisions to the slides.



## 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.

Table 1: Schedule of the Course

| Date | Lecture | Material | Assignment |
|---|---|---|---|
| 07/12 | 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. | |
| 07/12 | Joint and Conditional Probability. | Joint probability and conditional probability; Bayes' rule. Applications to college and income mobility. | HAGS 2.1, Lab 1 |
| 07/14 | Independence. Application of probability rules. | Definition of independence. Applications in evictions data. Applications in mobility data. | HAGS 2.1 |
| 07/16 | 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. | HAGS 2.1; Quiz 1; PSet 1 |
| 07/16 | 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. Distribution of profits conditional on formal and not formal. | HAGS 3.7; Quiz 1; PSet 1 |
| 07/19 | 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. | HAGS 2.2; HAGS 3.1; HAGS 3.2 |
| 07/21 | The Central Limit Theorem | Idea of simulation. The sample mean. Illustrating the Central Limit Theorem for sample means. See CLT in data visualization. Write random variable generating function, swap with another group, and calculate sample means. | HAGS 2.2; HAGS 3.1; HAGS 3.2, Lab 2 |
| 07/21 | 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. | HAGS 3.3; HAGS 3.4 |
| 07/23 | Randomized Control Trial | Reasons for Correlation? What is an RCT? Pros and Cons of RCTs. Examples of RCTs | Quiz 2; PSet 2 |

**Table 1 – continued from previous page**

| Date | Lecture | Material | Assignment |
|---|---|---|---|
| 07/26 | Review Session | | |
| 07/28 | Midterm exam | | Midterm exam |
| 07/30 | 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. | HAGS 4; HAGS 5 |
| 07/30 | Linear regression - Bivariate | R2 and goodness of fit. Running regressions in R and interpreting output. Specifying bivariate regressions. Functional form. | HAGS 6 |
| 08/02 | Linear regression - Multivariate | Estimating multiple regressions. Omitted variables bias. OVB formula and interpretation. Choosing regression specifications, good vs. bad control variables. | HAGS 7 |
| 08/04 | Lab 3: Multivariate Regression | Interpreting regression coefficients | |
| 08/06 | 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. | HAGS 10; HAGS 13.4 Quiz 3; PSet 3; Lab 4 |
| 08/06 | 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. | HAGS 13.4 Quiz 3; PSet 3; Lab 5 |
| 08/09 | Instrumental variables | Discussion of error term in OLS and identifying assumption of exogeneity. Introduction to IV, Wald Estimator and 2SLS; two-step and one-step implementations of 2SLS. | HAGS 12; Lab 5 |
| 08/11 | Final Review | | |
| 08/13 | Final Exam | | Final exam |