E392: Topics in Big Data

Spring 2018

Instructors:	Juan Carlos Escanciano	Stefan Weiergraeber
Email:	jescanci@iu.edu	sweiergr@iu.edu

Course Pages:

- Announcements and course materials will be posted on Canvas.

Lecture Time & Office Hours:

- Lecture: MW, 1:00pm–2:15pm, in WY 125
- Office hours: MW 2:30pm-4:00pm in WY 347 (Stefan Weiergraeber, Part 1) and WY201 (Juan Carlos Escanciano, Parts 2 and 3)

Objectives: The course consists of three parts. First, a practical part discussing how to work with data on your computer. Second, an introduction to popular tools from the field of statistical and machine learning. Third, an overview on recent advances on combining machine learning methods with economic models to conduct causal inference. In the first part, we will discuss the full workflow of data science from getting data, importing and cleaning data, visualizing data, and communicating the results of empirical analyses. Throughout the course we will use the software package R and economic data to illustrate the discussed concepts and methods. Example applications will include prediction and causal estimation of sales, wages, stock returns, credit defaults, etc.

Prerequisites: E370, E371 or equivalent

Main References: There are two main references below, one for the data management part and one for the statistical learning methods part. Both references are freely available in electronic form. The first reference (R4DS, henceforth) will be used for the first part of the course. The second reference (ISLR, henceforth) will be used for the second part. The third part, on causal analysis, will be based on recent research articles given below. Additional video material for parts two and three is given below.

- Hadley Wickham & Garrett Grolemund, *R for Data Science*, O'Reilly, 2016.
- James, G., Witten, D., Hastie & T., Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013

Software: For illustrating the methods and the empirical exercises you will need to have R, RStudio and several R libraries installed on your computer. All of these are free and open source. If necessary, we will use several tutorial sessions to introduce you to the fundamentals of the software.

page 1 of 4

Tentative Course Outline - Part 1:

Lecture 1: Introduction to R & RStudio ..... R4DS Ch. 2,4,6,8
Lecture 2: Data visualization ..... R4DS Ch. 3
Lecture 3: Data transformation ..... R4DS Ch. 5
Lecture 4: Exploratory data analysis ..... R4DS Ch. 7
Lecture 5: Data management ..... R4DS Ch. 10,11
Lecture 6: Tidy data & relational data ..... R4DS Ch. 12,13
Lecture 7: Programming in R ..... R4DS Ch. 19-21
Lecture 8: Communicating your results: R Markdown ..... R4DS Ch. 26-30

Tentative Course Outline - Part 2:

Lecture 9: Prediction & Causality ..... ISLR, Ch. 2; Reading: Varian (2014)
Lecture 10: Review of Linear Regression ..... ISLR, Ch. 3
Lecture 11: Lab: Linear Regression ..... ISLR, Ch. 3
Lecture 12: Linear Model Selection and Regularization ..... ISLR, Ch. 6
Lecture 13: *Lecture 12 continued* .....
Lecture 14: Lab: Linear Model Selection and Regularization ..... ISLR, Ch. 6
Lecture 15: Tree-based Methods ..... ISLR, Ch. 8
Lecture 16: Lab: Tree-based Methods ..... ISLR, Ch. 8
Lecture 17: Moving beyond linearity ..... ISLR, Ch. 7
Lecture 18: Lab: Moving beyond linearity ..... ISLR, Ch. 7

Tentative Course Outline - Part 3:

Lecture 19: Rubin Causal Model ..... Athey (2017) and Athey et al. (2017)
Lecture 20: Inference with many Controls ..... Chernozhukov et al. (2017)
Lecture 21: Instrumental Variable Methods .....
Lecture 22: Example: Demand Estimation ..... Bajari et al. (2015)

Grading Policy: Homeworks (20%), empirical group project (40%), final exam (40%). There will be approximately 8 problem sets throughout the semester. Please work on and hand in each problem set in groups of at most 2 students. Each problem set will consist of several questions. You are required to hand in all questions, but we will randomly select only 2 questions of each problem set to grade.

Empirical Group Project: At the end of the semester students in groups of at most two will have to hand in an empirical project in which you gather some data, run some analysis and present the results in a short report. You are free to choose the topic of your project as long as it relates to some of the topics discussed in class. See below for important information on dates. The project cannot exceed the limit of 20 pages.

Important Dates:

Project Proposal: Two page proposal due on **March 19th**, 2018.
Final Project: The project will be due on **April 23rd**, 2018.
Final Exam: **May 2nd**, 2018

page 2 of 4

References:

- Athey, S. (2017): "Beyond Prediction: Using Big Data for Policy Problems," Science.
- Athey, S., G. Imbens, T. Pham, and S. Wager (2017): "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges," American Economic Review, forthcoming.
- Bajari, P., Nekipelov, D., Ryan, S. and M. Yang (2015): "Machine Learning Methods for Demand Estimation", American Economic Review: P&P, 105(5).
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C.Hansen, and W. Newey, (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects," American Economic Review: P&P, 107(5), 261-65.
- Einav, L., and Levin, J. (2014): "Economics in the Age of Big Data," Science, 346.
- Hadley Wickham & Garrett Grolemund, *R for Data Science*, O'Reilly, 2016. Available at http://r4ds.had.co.nz/index.html.
- James, G., Witten, D., Hastie & T., Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013. Available at http://www-bcf.usc.edu/ gareth/ISL/.
- Mullainathan, S. and J. Spiess (2017): "Machine learning: an applied econometric approach," Journal of Economic Perspectives, 31, 87-106.
- Varian, H.R. (2014): "Big Data: New Tricks for Econometrics", Journal of Economic Perspectives, 28(2), 3-28.

Video and Slides Material:

- Hastie and Tibshirani: https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
- NBER (by Athey and Imbens): http://www.nber.org/econometrics_minicourse_2015/

Policies:

- **Academic Integrity:** As a student at IU, you are expected to adhere to the standards and policies detailed in the Code of Student Rights, Responsibilities, and Conduct. When you submit an assignment with your name on it, you are signifying that the work contained therein is yours, unless otherwise cited or referenced. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged. If you are unsure about the expectations for completing an assignment or taking a test or exam, be sure to seek clarification beforehand. All suspected violations of the Code will be handled according to University policies. Sanctions for academic misconduct may include a failing grade on the assignment, reduction in your final course grade, a failing grade in the course, among other possibilities, and must include a report to the Dean of Students, who may impose additional disciplinary sanctions.

page 3 of 4

- **Special circumstances:** Students requiring any type of special classroom/testing accommodation for a disability, religious belief, scheduling conflict, or other impairment that might affect his or her successful completion of this course must personally present the requested remedy or other adjustment in written form (signed and dated) to the instructor, i.e. supporting memorandum of accommodation from the Office of Disabilities Services for Students. Requests for accommodations must be received and authorized by the instructor in written form no less than two weeks in advance of need. No accommodation should be assumed unless so authorized. In the event of needs identified later in the course, or for which an adjustment cannot be made on a timely basis, a grade of "I", Incomplete, for the course will be given to accommodate the unanticipated request.

- **Exam absences:** In the event of a catastrophic (and documented) occurrence which necessitates an absence from a scheduled exam, the student should immediately seek the instructor's *permission to miss an exam.* If approval is granted, the weights of the student's scores for the other exams will be re-adjusted proportionately, so as to make up for the missed exam. If completed documentation is not presented within one week after a missed exam, or if no *permission to miss a exam* has been obtained prior to the exam date, the missed exam will received a score of zero points.

page 4 of 4