Economics 484

Econometrics and Data Science

(Machine Learning for Econometricians)

Department of Economics
University of Washington-Seattle
Spring 2015

Professor Gregory M. Duncan
Affiliate Professor of Economics
Chief Economist and Statistician, Amazon.com

Lecture: MW 5:30-7:20 pm DEM 104
Section: W 7:30-8:20 pm SAV 117
Pasita Chaijaroen
Graduate Assistant

1) Weeks 1 and 2 (Review)
   Read Chapters 1-3 before class. Chapter 3 will be skipped as it is a summary of part of my Econ 482 course. Compare with Ch 4-7 in Wooldridge.
   a) Review of matrices Wooldridge Appendix D
   b) Review of linear model in matrices Wooldridge Appendix E
      i) use of lm()
   c) The Experimental Model
   d) Quasi-experiments
   e) Structural Models
   f) Threats to validity
      (1) Endogeneity
      (2) Selectivity
   g) Methods for analyzing observational data
   h) Elements of R-Programming (in Sections 1 and 2)
      i) use of lm()
      ii) use of sandwich()
2) Weeks 3-8 Supervised Learning
   a) Nonlinear estimation
      i) Review of central limit theorem and the law of large numbers. Wooldridge Appendix C.3
         (1) Slutsky's Calculus
      ii) Maximum likelihood and GMM Wooldridge Appendix C.4
         (1) Numerical optimization in one easy lesson
      iii) Instrumental Variables Wooldridge Ch 15
         (1) STATA and R
      iv) Classification and logistic regression Ch 4
   b) Cross-Validation and Bootstrapping Ch 5
   c) Variable Selection/Feature Engineering Ch 6

(1) the Information Criteria
(2) the LASSO
(3) elastic net
(4) Shrinkage Methodss

d) Non-parametrics Ch 7-8
i) Trees
ii) Splines
iii) The curse of dimensionality
iv) Large data approach
(1) Boosting
(a) ADABoost
(b) ADABoost.L
(2) Bagging
(3) Random Forests

3) Weeks 9-10
a) The Structural Modeling Approach Pearl (2009), Haavelmo(1944), Wooldridge Ch 15-16
i) The probability method in Econometrics
(1) The identification Problem
(2) Recursive Models
(3) General Models
ii) The Structural Causal Modeling Approach of Computer Science
(1) Pearl's do() Calculus
b) Instrumental Variables Methods
i) Quasi-experiments revisited
ii) Selectivity

4) Unsupervised Learning (if there is time)
a) Clustering Ch 10.3

Texts and Other Material:

Required Text:

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) *An Introduction to Statistical Learning: with Applications in R* (Springer Texts in Statistics), Springer.

Other Material:

Angrist, Joshua D. and Jörn-Steffen Pischke, (2008). *Mostly Harmless Econometrics: An Empiricist's Companion,* Princeton University Press

Berk, Richard A., (2009) *Statistical Learning from a Regression Perspective,* Springer

Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations". *Econometrica*, Vol. 11, 1–12.

Haavelmo, T. (1944). "The Probability Approach in Econometrics" *Econometrica*, Vol. 12, Supplement, iii-115

Heckman, James J. and Rodrigo Pinto (2012) *Causal Analysis after Haavelmo: Definitions and a Unifed Analysis of Identification of Recursive Causal Models*, Causal Inference in the Social Sciences, University of Michigan

Heckman James J. (2010) "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy", *Journal of Economic Literature*, Vol. 48, No. 2

Paul W. Holland (1986) "Statistics and Causal Inference", *Journal of the American Statistical Association* Vol. 81, No. 396, pp. 945-960

Morgan, Stephen L. and Christopher Winship, (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research,* Cambridge University Press

Pearl, Judea (2009) "Causal inference in statistics: An overview" *Statistics Surveys* Vol. 3 96–146

Wooldridge, Jeffrey M. (2012) *Introductory Econometrics: A Modern Approach*, Cengage Learning; 5th edition

This course is an advanced continuation of Economics 482 and 483. It assumes a good background in regression at the level of the Wooldridge text above. It will cover topics such as Simultaneous Equations Modeling (Structural Modeling, Instrumental Variables), Non-linear modeling (non-linear regression, logit, probit, maximum likelihood, with a brief, heuristic, introduction to Generalized Method of Moments), Variable Selection using the LASSO, and Modern Non-parametric Modeling from a Machine Learning Perspective (Regression and Classification Trees, Bagging, Boosting, and Random Forests). The course is decidedly hands on emphasizing interpretation, not formal proofs. That said it, uses math and stat skills and concepts without apology or review. The course is ideal for double majors in Economics with Math/Stat or Computer Science as the other major or graduate students in Economics, Business, Public Policy or the other social sciences.

Prerequisites: Econ 482, Econ 483, Math 126 and familiarity with matrices and basic matrix operations (structure, transpose, inverse, multiplication). Knowledge of one major statistical program (SAS, STATA, SPSS) and some familiarity with R. Those who took my Econ 482 should have the background.

Learning Goals: By the end of the course the students will be able to use R and STATA to analyze large datasets using a variety of new tools taught in the course. These new tools include instrumental variables, non-linear estimation, the LASSO for variable selection and Random Forests. Particular emphasis will be put on instrumental variables estimation in the Roy Model (average treatment effects), binary and multinomial logistic regression. They will understand which tools are called for by the different structures of the data and the underlying reason for analysis. So for example, for a label response variable, a logistic type regression model might be best for interpretation, but random forests might be best for prediction.

The overall learning goal include providing sufficient background in machine-learning, as applied to economic problems, so as to make the students able to get jobs as research assistants and analysts at organizations using or interested in using so called "analytics" and "big data" methods. Such places would include major consulting companies (e.g. NERA, Deloitte, Brattle), major technology companies (e.g. Amazon, Google, Tesla), major retailers (e.g. Nike, The Gap, Nordstrom) or government agencies (e.g. FTC, DOJ, IMF). The sufficient background alluded to include the ability to setup, run and interpret the output of the methods learned in R and or STATA.

Grading: 30% homework, which will be primarily computer oriented. 70% final, which will test interpretation of computer output, set up of analysis and tool and model identification.

Why is this course needed?

Analytics and data science are becoming important at many firms. Currently only advanced degree holders are competent to do even the simplest big data tasks, even though these are conceptually no more difficult than standard econometrics as taught in the best undergraduate programs. As a consequence many firms simply do not bother to try these methods that could help them or they use primitive spreadsheet methods, or, at best regression methods that are inappropriate for large data.

This course will do two things, first provide sufficient background in machine-learning, as applied to economic problems, so as to make the students able to get jobs as research assistants and analysts at organizations using or interested in using so called "analytics" and "big data" methods. Such places would include major consulting companies (e.g. NERA, Deloitte, Brattle), major technology companies (e.g. Amazon, Google, Tesla), major retailers (e.g. Nike, The Gap, Nordstrom) or government agencies (e.g. FTC, DOJ, IMF).

Second, prepare those going on to graduate school with a broader perspective of what econometrics is changing into.

Course Description: (for Catalog)

Advanced continuation of Economics 482 and 483. Covers traditional topics: Structural Modeling, non-linear and logistic regression, the LASSO, and non-traditional topics: Regression and Classification Trees, Bagging, Boosting, and Random Forests). The course is computer based, using the R language, emphasizing interpretation, not formal proofs.