Economics 484

Econometrics and Data Science

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

TA Pasita Chaijaroen <*pasita@uw.edu*>

1) Week 1
    a) The Experimental Model
    b) Quasi-experiments
    c) Structural Models
    d) Threats to validity
        (1) Endogeneity
        (2) Selectivity
    e) Methods for analyzing observational data
    f) Elements of R-Programming (in Sections 1 and 2)
2) Weeks 2-7
    a) Review and extension of regression models.
        i) Approximation by piecewise constants
        ii) Nearest neighbor ideas
    b) Variable Selection
        (1) the LASSO
        (2) elastic net
        (3) Shrinkage Models
    c) Nonlinear estimation
        i) Review of central limit theorem and the law of large numbers.
            (1) Slutsky's Calculus
        ii) Maximum likelihood and GMM
            (1) Classification and logistic regression
                (a) Multinomial logistic regression
            (2) STATA and R
        iii) Non-parametrics
            (1) Trees
            (2) Splines
            (3) The curse of dimensionality
            (4) Large data approach
                (a) Boosting
                    (i) ADABoost

(ii) ADABoost.L
(b) Bagging
(c) Random Forests

3) Weeks 8-10
   a) The Structural Modeling Approach
      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

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

*Belloni, Alexandre and Victor Chernozhukov (2013) Least squares after model selection in high-dimensional sparse models, *Bernoulli* 19(2), 521–547 DOI: 10.3150/11-BEJ410

Peter Bickel and Kjell Doksum, (2015) *Mathematical Statistics: Basic Ideas and Selected Topics*, Volume I, Second Edition Chapman and Hall/CRC

Peter Bickel and Kjell Doksum, (2015) *Mathematical Statistics: Basic Ideas and Selected Topics*, Volume II Chapman and Hall/CRC

*David Donoho, 2015, *50 years of Data Science*, Tukey Centennial Workshop, http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations". *Econometrica*, Vol. 11, 1–12. http://www.jstor.org/stable/1905714

Haavelmo, T. (1944). "The Probability Approach in Econometrics" *Econometrica*, Vol. 12, Supplement, iii-115 http://www.jstor.org/stable/1906935

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

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

Pearl, Judea (2014) "Trygve Haavelmo and the Emergence of Causal Calculus" Econometric Theory, Special Issue on Haavelmo Centennial. http://ftp.cs.ucla.edu/pub/stat_ser/r391.pdf

The Belloni and Chernozhukov article and the Donoho article are required reading. Everything from Haavelmo on is about the area of causality, as is the Angrist book, an area of heated controversy. The Bickel and Doksum books are very hard, but they are at the level of the kind of statistics you would need in grad school. The chapter in Volume II on Machine learning is particularly good though very, very dense. These books are not needed for this course.

Recommended Background Text:

Jeffrey M. Wooldridge, (2016) *Introductory Econometrics: A Modern Approach*, Cengage Learning; 6th edition

Joseph Adler, (2012) *R in a Nutshell* (In a Nutshell (O'Reilly)) O'Reilly Media

Hadley Wickham, (2014) *Advanced R* Chapman and Hall/CRC

The R books are useful but there are free sites all over the web.

The Wooldridge book has econometrics at the level I expect for people taking this course, I will often refer to it and will assign some readings.

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, 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 includes 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.