Modern Methods for Applied Regression
PPHA41430/ECONXXXXX

September 16, 2024

Instructor: Guillaume A. Pouliot

Teaching Assistant: Jake Nicoll

Lecture room and time: Section 1: T/Th at 3:30-4:50pm Keller 0007; Section 2: 5:00 - 6:20pm

Section room and time: TBD

Course Description

The course aims to take a mature and modern take on econometrics and regression methods more specifically. The starting point is that standard reduced form methods such as ordinary least-squares (OLS) are almost always misspecified in practice. For instance, textbooks and research material often refer to OLS as estimating the conditional expectation, but this is only true if the conditional expectation is linear, which it almost never is in practice.

From that point, two avenues can be taken. A first avenue is to ask whether the estimator is still meaningful. In particular, we can ask whether OLS still estimates, in some sense, a best approximation to the (nonlinear) conditional expectation, even though it cannot estimate the conditional expectation itself. We will see that this is sometimes the case, and the course will give a careful treatment of estimators with such a best approximation property.

A second avenue is to ask whether a more flexible estimator may in fact recover the desired estimand. This is where machine learning methods "come to the rescue". In the case of reduced form methods, recent development in double/debiased machine learning (DML) have allowed us to tackle such modeling issues using modern machine learning methods taken right off the shelf. We will cover deep learning methods "from scratch" and use them to recover correct specification in reduced form estimation.

Regression analysis requires inference. We cover three fascinating topics in inference. First, we cover Markov chain Monte Carlo Methods (MCMC), treating a host of methods using the Metropolis-Hastings algorithm as a unifying concept and algorithm. Second, we explore gains in robustness from using randomization-based inference methods –on observational data– to carry out inference with reduced form regression methods such as linear regression. Finally, we consider approaches to producing confidence intervals which may be considered simultaneously without invalidating their coverage guarantees.

Finally, we cover reinforcement learning, and pay special attention to connections with econometric and statistical questions.

Background and Goals

Background in linear algebra, probability, statistics and econometrics. Students who have successfully completed a first undergraduate course in econometrics are considered to have sufficient background.

The goal of the course is to provide students with a firm conceptual understanding of the reduced form regression methods and statistical inference, as well as practical experience with a selection of corresponding statistical and econometric methods.

Grades

Students will be evaluated based on participation, weekly problem sets, and a final presentation.

- Participation: 15%
- Weekly problem set: 70%
- Final presentation: 15%

For the final presentation, students must collect or create a dataset and carry out and present a regression analysis of the data.

Timeline

Topics in estimation

Week 1 Linear Regression for grown ups

Ordinary linear regression (OLS) is the workhorse of reduced form regression analysis. We study its meaning when the conditional expectation, the oft cited estimand of linear regression, it itself nonlinear and thus cannot possibly be estimated with a linear model.

readings: "Linear Regression for Grown Ups" handout; Mostly Harmless Econometrics, Chapter 3 up to and including 3.1.2.; Gary Chamberlain Lecture Notes, Lecture note 1

Week 2 Quantile regression

While OLS allows a characterization or approximation of the conditional expectation, quantile regression allows for the study of the entire conditional distribution. We study quantile regression and pay special attention to misspecification, inference, and computation.

readings: "Modern Introduction to Quantile Regression" handout

Week 3 Instrumental variables

We revisit two-stage least-squares in light of lessons learned. In particular, we ask where best approximation properties obtain and where they don't, anticipating the next section in which machine learning methods will "come to the rescue". We revisit classical topics such as heterogeneous treatment effects, weak instruments and many instruments.

readings: "Many and Weak IV" handout, handbook chapter draft

Topics at the intersection of estimation econometrics and machine learning

Week 4 Backgrounder on machine learning

Some of the specification problems raised in the previous section of the course may be addressed using modern machine learning methods. We revisit those and try our hand at using them in practice. We pay particular attention to neural networks, random forests, and support vector machines.

readings: An Introduction to Statistical Learning with Applications in R, Chapters 8-10.

Week 5 Machine learning solutions for reduced form regression problems

We revisit high-dimensional linear regression and instrumental variables linear regression and extend their applicability using modern machine learning methods. We pay particular attention to the debiased/double machine learning (DML) approach.

readings: TBD.

Topics in inference

Week 6 Markov chain Monte Carlo

We give a primer on inference using MCMC methodology. We use Metropolis-Hastings as a unifying concept and algorithm.

readings: "MCMC" handout

Week 7 Randomization-based inference

Modern randomization based inference offers an alternative toolkit for more robust inference at moderate costs in power. It furthermore offers a principled approach to "placebo" tests, allowing for instance the use of permutation tests that produce valid p-values even on observational, non-experimental data.

readings: "Inference" handout

Week 8 Simultaneously valid inference

We first stress the point that every time we consider the coverage probabilities of two ore more confidence intervals simultaneously, we are implicitly testing multiple hypotheses, meaning that the marginal coverage statements do not correspond to the implied –familywise error rate– coverage statement. We explore classical and modern approaches to multiple hypothesis testing, and consider applications ranging from the two confidence interval example to the testing of millions of hypotheses in big data applications.

readings: "Simultaneously valid inference" handout

Dynamic programming

Week 9 Reinforcement learning

Many challenging tasks in economics and in the "tech" industry involve solving a sequence of problems where the conditions of one problem are influenced by actions taken in the previous problem. Modern methods from machine learning and econometrics have allowed us to solve such dynamic programming problems at scale and to carry out inference on key statistics and coefficients, in spite of the sophisticated structure of the problem. We explore these tools and some of their applications,

Student presentations

Week 10 Student presentations

Time and class size allowing, we will have short student presentations.

References

Angrist, J.D. and Pischke, J.S., 2009. Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Chamberlain, Gary. Lecture Notes.
https://github.com/paulgp/GaryChamberlainLectureNotes/tree/main/RawLectures