ECO 517a Fall 2020

Lecture 1: Approaches to data analysis

Christopher A. Sims
Princeton University
sims@princeton.edu

July 28, 2020

Why start with this?

- There are new ideas and new approaches in econmetrics, some complementing each other, some competing. By briefly describing them and providing references, I'm hoping some of you will become inspired.

- In this introductory course we can't give each of these approaches a full exposition. The course will take its own approach, taking elements from those described in these slides. You should be aware that this course takes a point of view that might be controversial.

Traditional econometrics

- “Theorists” present models (in principle, probability models) that contain unknown “parameters”.

- “Econometricians” estimate those parameters or test whether they have values predicted by theory.

- Economic science advances in a process of generating or advancing theories with “testable implications”, rejecting those that econometricians determine are in conflict with the data.

Traditional econometrics

- "Theorists" present models (in principle, probability models) that contain unknown "parameters".
- "Econometricians" estimate those parameters or test whether they have values predicted by theory.
- Economic science advances in a process of generating or advancing theories with "testable implications", rejecting those that econometricians determine are in conflict with the data.
- No measurement without theory!

Machine Learning

- “Subject matter specialist” presents gigabytes or terabytes of data that he suspects must contain useful knowledge.

- A computer scientist applies sophisticated algorithms that find patterns in the data, or show that the data can be used to predict something.

- Useful things are learned, while an econometrician would barely have begun figuring out what theoretical model he or she should bring to the data.

Causal inference

- This approach arose out of work on clinical trials and policy experiments where a “treatment” is applied $(T_i = 1)$ or not applied $(T_i = 0)$ to large numbers of individuals i.

- It brings to the fore concerns about heterogeneity of treatment effects across individuals and about biases in the selection of who gets the treatment.

- It de-emphasizes explicit probability modeling, instead making “weak” assumptions that justify “large sample” approximate measures of uncertainty about estimators.

Bayesian inference

- Uses explicit probability models to arrive at probability distributions for unknown parameters or for unobserved variables.

- This approach has always had a following, but until the last couple of decades it was seen as impractically demanding computationally. New approaches based on computer simulation have made it much easier to implement.

Relations among these approaches

- Though each approach can be the basis for criticizing the others, there is a reason each has a following and enthusiastic proponents.

Machine learning as measurement without theory

- Machine learning, particularly the great recent success of “deep neural nets”, has managed to sidestep the effects of its lack explicit measures of uncertainty about results.

- In most economic applications there is considerable uncertainty about which model, or which version of a model “fits best”, and measuring that uncertainty is important to making results useful.

- In image or voice recognition, where deep neural nets have had great success, success can be measured simply by whether the computer has matched the abilities of a human performing the same task.

- There are no parameters within a model whose value is of independent interest, and in fact often it is clear that the algorithm used finds only “a good-enough model”, not even an attempt to find the best possible one.

Choosing model complexity

- Machine learning data analysis always has model complexity tuning parameters (e.g., the number of groups k in k-means, the number of layers in a neural net), and choosing them is part of the data analyst's task.

- Traditional econometric approaches, and the treatment-effect causal inference literature, tend to discuss how to estimate a model with a given finite number of parameters — though in fact applied researchers often do consider an array of models with varying complexity.

- Bayesian inference provides a way to go beyond informal rules of thumb in choosing such complexity levels.

The approach of this course

- Traditional econometrics, mostly from a Bayesian perspective, and following machine learning in treating model complexity choice in most applications.

- I'll try to have an exercise ready each week that connects our discussion of models and statistical theory to data analysis.

This week: k-means

- The exercise this week has you apply a “machine learning” algorithm that tries to extract patterns from data.

- You only need to know how to feed the data to the computer algorithm to do it. There is no complicated theory to show the algorithm is optimal or to attach uncertainty measures to the result.

- You’ll look at (some of) the data that Angrist and Krueger used to study the returns to education, based on a sample from the US census containing data on over 300,000 men.

- We’ll come back to these data with other approaches, but here you will use the k-means algorithm (by calling `kmeans()` in R, unless you want to use other software) to allocate the individuals in the data set to a small number of groups.

What k-means does

- You give it data in an $N \times m$ array X, with each row X_i representing values of some variables for an individual i.

- You specify how many groups you want: k.

- The algorithm generates k m-dimensional mean vectors μ_j and allocates each individual to a group $j(i)$.

- It tries to do this in such a way that the sum of squared distances between observations and group means, $\sum_i \|X_i - \mu_{j(i)}\|^2$ is minimized.

Criticisms of k-means

- If the data are in fact drawn from groups, this algorithm is not the optimal way to discover the groups.
- If we knew which group each individual belonged to, it would indeed be optimal to find the μ_j's by minimizing the sum of squared deviations from group centers.
- If we knew the group centers and needed to allocate each individual to a group, allocating each X_i to the group with the nearest μ_i could make sense.
- But the full algorithm does not make sense.

A simple example

- Suppose m is 1, so that each X_i is a single number. And suppose that the X_i's are (0, .1, .5, .9, 1).

- If we give these data to k-means, with $k = 2$, it will either say the first three observations are group 1, and the last 2 group 2, or else the first 2 are group 1 and the last 3 are group 2. We'll have $\mu = (.2, .95)$ or $\mu = (.05, .8)$. Both give the same sum of squared deviations.

- But in reality we must be very uncertain about which group the .5 observation belongs to.

- Also, changing the .5 observation to .51 or .49 makes the solution unique, but the tiny change in the one observation causes a jump in the estimated μ vector.

So why use it?

- It's easy to use and it's easy to understand what it does.
- The results are sometimes useful, for example in discovering how much or what type of heterogeneity is present in a sample.
- There are extensions of k-means, some derived from probability models, that are better, but they are usually computationally more demanding and often give results that are not too different.
- The problem of observations whose allocation to groups is uncertain is not the only, and not the worst, problem with k-means. See the running discussion of k-means through several chapters of the McKay book on the reading list.