Machine Learning and Economics

Dweepobotee Brahma
IIT Jodhpur

NISER Bhubaneshwar , 16th October, 2023

Setting the Stage

- ML and Economics (and Econometrics) are two apparently divergent fields yet have a lot in common.
- They differ in their philosophy of empirical modelling and goals.
- Yet modern advances in computation and Big Data have brought these two fields closer again.

SETTING THE STAGE

Economics is the study of choices we make under scarcity.

SETTING THE STAGE

Economics is the study of choices we make under scarcity.

All economics questions arise because we have unlimited wants but limited resources.

SETTING THE STAGE

Economics is the study of choices we make under scarcity.

All economics questions arise because we have unlimited wants but limited resources.

This leads to scarcity and we need to make choices.

SETTING THE STAGE

Economics is the study of choices we make under scarcity.

All economics questions arise because we have unlimited wants but limited resources.

This leads to scarcity and we need to make choices.

Note: Scarcity can be in anything. E.g. Money, personnel, physical infrastructure, time, attention.

SETTING THE STAGE

- Economic Theory helps to identify *deterministic* relationships between variables.
  - Theory tells us about the direction of change in one varible in relation to the change in another.
  - Theory is also a simplified representation of reality.
- In practice relationships between variables are *stochastic*, not *deterministic*.
  - Incorporating stochastic elements transforms the theory to one from an *exact* statement to that of a probabilistic statement using *expected* outcomes.

SETTING THE STAGE

Econometrics picks up where economic theory leaves off.

- Econometric models *start by* typically taking a statement from economic theory and proceeds to test this statement using data.
- Econometrics requires theory $\rightarrow$ data $\rightarrow$ statistical tools

ECONOMETRICS

The term 'Econometrics' was coined by Ragnar Frisch, 1969 Economics Nobel Prize co-winner.

"*Econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics.Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics*."

Frisch (1933), Econometrica, vol 1, pgs 1-2

ECONOMETRICS

Sounds familiar??

ECONOMETRICS

Sounds familiar??

Similar to how we define 'data science'.

ECONOMETRICS

Sounds familiar??

Similar to how we define 'data science'.

Emphasis on incorporating domain knowledge (from economic theory) to build models.

ECONOMETRICS

Sounds familiar??

Similar to how we define 'data science'.

Emphasis on incorporating domain knowledge (from economic theory) to build models.

Developing statistical models for answering economics (or even wider social science) questions has put econometrics on a specific trajectory.

ECONOMETRICS

Thus econometrics has traditionally focussed on questions related to

- marginal effect
  - E.g. How many extra children can be vaccinated if an additional one lakh rupee was allocated?
  - The '**how much**' question.
- causal inference
  - E.g. Does advertising cause an increase in sales? How Much?
  - The **'Why'** question.
  - E.g. What is the size of the effect of advertising on Firm's sales?
  - The **How much** question

Machine Learning

In contrast, Machine Learning has been about prediction.

MACHINE LEARNING

In contrast, Machine Learning has been about prediction.

The starting point of most empirical modelling is the linear regression model.

$$Y = \alpha + \beta X + u \tag{1}$$

Econometrics has been all about $\hat{\beta}$.

MACHINE LEARNING

In contrast, Machine Learning has been about prediction.

The starting point of most empirical modelling is the linear regression model.

$$Y = \alpha + \beta X + u \tag{1}$$

Econometrics has been all about $\hat{\beta}$.

Machine Learning has focussed on $\hat{Y}$.

Machine Learning in Economics

So how does Economics use Machine Learning?
There are three commonly used ways (so far) in empirical research.

- Use ML tools to harvest non traditional data (usually Big Data) - Big Data and Economics.
- Pure prediction problems - predict health/labour/education events/outcomes
- Combine inference with prediction - causal ML (economists' contribution to the ML literature)

Big Data in Economics

Problem: Estimates of wealth, GDP, economic activity have traditionally been generated at state or district level using large scale surveys or census (expensive to conduct) at long intervals (every few years).

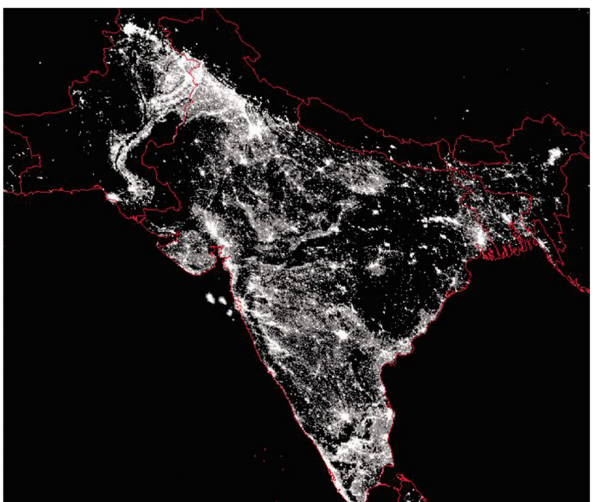
Big Data in Economics

Problem: Estimates of wealth, GDP, economic activity have traditionally been generated at state or district level using large scale surveys or census (expensive to conduct) at long intervals (every few years).

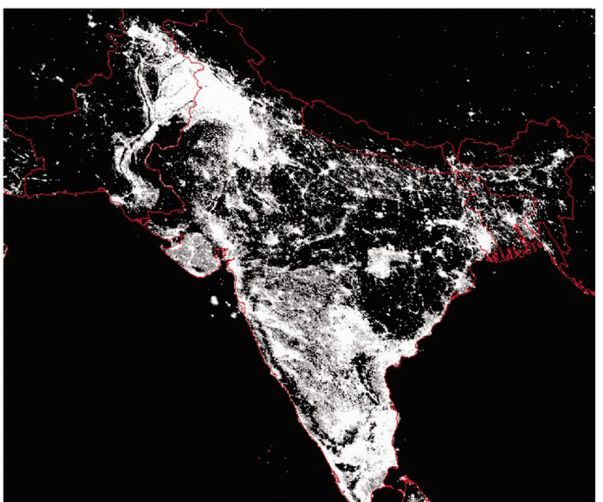
Solution: Used satellite imagery to convert luminosity from artificial light at night to convert into a measure of economic activity at a spatially granular level.

- Nightlights data has been used to generate more local estimates of economic activity and poverty. (Pinkovsky and Sala-i-Martin(2016))
- Used to look at the effect of Covid lockdown on economic activity. (Beyer et al. 2021)

Satellite images of South Asia by night



South Asia in 1994



South Asia in 2010

Images are taken from Maxim Pinkovskiy and Xavier Sala-i-Martin (2016) – *Lights, Camera … Income! Illuminating the National Accounts-Household Surveys Debate*. The Quarterly Journal of Economics

Big Data in Economics

Problem: Certain aspects of economic and industrial policy cannot be quantified i.e. information about tone and stance of a government cannot be measured well through traditional data.

Big Data in Economics

Problem: Certain aspects of economic and industrial policy cannot be quantified i.e. information about tone and stance of a government cannot be measured well through traditional data.

Solution: Using government bulletins, memos, press releases to construct text based indices.

- Juhasz et al.(2022) develop a cross-country text based industrial policy index and take it to a global database of economic policies.
- Ash et al.(2021) use a dataset of 80 million cases in the Indian judicial system from 2010-2018 to investigate in-group bias by gender and religion in the lower courts of Indian judiciary.

Prediction Problems in Economics

Standard econometric models help you identify the **average effect** of an intervention.

Prediction Problems in Economics

Standard econometric models help you identify the **average effect** of an intervention.

Machine Learning can be used to identify **who** to target interventions to - better allocation of resources.

Prediction Problems in Economics

Standard econometric models help you identify the **average effect** of an intervention.

Machine Learning can be used to identify **who** to target interventions to - better allocation of resources.

It can answer the **who** and the **where**?

Prediction Problems in Economics

- A public-sector resource allocation problem is the question of how a city should allocate building inspectors optimally to minimize safety or health violations.
  - New York City's Firecast algorithm allocates fire inspectors according to the predicted probability of a violation being detected upon inspection.
  - Similar to Glaeser et al.system for allocating health inspectors to restaurants in Boston,
  - Preliminary estimates showing a 30 to 50% improvement in the number of violations found per inspection.

Prediction Problems

- Lentz (2019) predict food security status in Malawi by incorporating granular market data, remotely-sensed rainfall and geographic data, and demographic characteristics. - the **where**?
- Jean et al. (2016) combine satellite data with ML to predict poverty. - the **where**?
- Brahma and Mukherjee (2022)"Early Warning Signs: Targeting Neonatal and Infant Mortality using Machine Learning" in Applied Economics.- the **who**?
- Jayachandran et al. (2021) use ML to identify the best survey closed-ended questions to predict an agency score measured through qualitative interviews. - the **which**?

CAUSALITY IN ECONOMETRICS

Causality has been the central focus in econometrics following the Roy-Rubin 'potential outcomes' framework - popularly called 'what-if' scenarios. Focus is on causal inference i.e. measuring **the effect of a cause**, not **causes of an effect**.

Causality in Econometrics

Causality has been the central focus in econometrics following the Roy-Rubin 'potential outcomes' framework - popularly called 'what-if' scenarios. Focus is on causal inference i.e. measuring **the effect of a cause**, not **causes of an effect**.

Despite the recent advances in field-experiments, most empirical research in economics uses observational data. Additionally, economics often works with non-iid data especially panel data/longitudinal data.

Causality in Econometrics

Econometrics has seen the '*credibility revolution*' in the last few decades - through the development of many causal inference techniques for observational data.

- randomized controlled trials/lab-in-field experiment (Duflo, Kremer and Banerjee, Nobel Prize 2019)
- natural experiments/quasi-experiments (Card, Angrist and Imbens, Nobel Prize 2021)
- difference-in-difference, event-study techniques, two-way fixed effects
- regression discontinuity, interrupted time series
- instrumental variables
- propensity score matching
- synthetic control method

Causality in Econometrics

The fundamental problem of causal inference is that for each individual, we only get to observe one of the two potential outcomes!

Causality in Econometrics

The fundamental problem of causal inference is that for each individual, we only get to observe one of the two potential outcomes!

In other words, to get the causal effect, we need to observe both realities where an event happened and the alternate reality where the event didn't happen.

Causality in Econometrics

The fundamental problem of causal inference is that for each individual, we only get to observe one of the two potential outcomes!

In other words, to get the causal effect, we need to observe both realities where an event happened and the alternate reality where the event didn't happen.

In other words, this approach treats causal inference as a problem of missing data.

Causality in Econometrics

The fundamental problem of causal inference is that for each individual, we only get to observe one of the two potential outcomes!

In other words, to get the causal effect, we need to observe both realities where an event happened and the alternate reality where the event didn't happen.

In other words, this approach treats causal inference as a problem of missing data.

All methods of causal inference under the potential outcomes framework boil down to finding different ways to *uncover* this missing data - i.e. *predicting* the counterfactual.

Causality in Econometrics

The fundamental problem of causal inference is that for each individual, we only get to observe one of the two potential outcomes!

In other words, to get the causal effect, we need to observe both realities where an event happened and the alternate reality where the event didn't happen.

In other words, this approach treats causal inference as a problem of missing data.

All methods of causal inference under the potential outcomes framework boil down to finding different ways to *uncover* this missing data - i.e. *predicting* the counterfactual.

Econometrics meets Machine Learning!

Choosing the right nudge

- Child immunization is one of the most cost effect ways to reduce child mortality and morbidity.Yet take-up is low across the developing world.
- Banerjee et. al. (2021) ran a field experiment with many different nudges. E.g. mall incentives in cash or kind,symbolic social rewards,SMS reminders,and the use of influential individuals in society or in the social network as “ambassadors.”
- All of these strategies may work in various contexts

Combining Econometrics with Machine Learning

- There may be different dosage variants within each strategy. Policies may work best in tandem or counteract each other. Want to know **which combination** of nudges works best.
- We need to know which of these nudges is the most effective (i.e., leads to the largest increase in immunization), and the most cost-effective (i.e., leads to the largest increase in immunization per rupee spent). **How much?**
- Out of 75 possible intervention combinations of the nudges Banerjee et al. (2021) use causal-ML to select the most effective and most cost effective nudges and conduct post-selection inference to get the effect size of these interventions (econometrics).
- Find that SMS reminders and "information hubs" (influential individuals in the social network) are most effective.

Combining Econometrics with Machine Learning

Problem: A policy design question is often to understand how different subpopulations respond (treatment effect heterogeneity) to a particular intervention and what these subpopulations are.

- Standard econometric models will generate the average effect of an intervention.
- Can expand the analysis using interaction terms.
- But *ad hoc* searches for particularly responsive subgroups may mistake noise for a true treatment effect.

Combining Econometrics with Machine Learning

- Athey and Wager (2019) develop the statistical theory for 'causal forest' (combine random forest with causal inference).
- Tweak the random forest algorithm to identify subgroups who have similar treatment effects of the intervention. Instead of prediction error, causal tree/forest uses a treatment-control difference of mean outcomes to predict conditional average treatment effect (CATE) in each leaf.
- Heller and Davis (2017) use data from a large intervention providing summer jobs to disadvantaged youth and apply causal forest to identify two specific subgroups on whom this intervention was most effective.

Combining Econometrics with Machine Learning

Problem: One goal of policy research is often to identify which policy – out of a range of possible design or implementation options – will have the greatest impact on the outcome in question. RCTs require painstaking and costly data collection with potentially long periods of observation.

Combining Econometrics with Machine Learning

Problem: One goal of policy research is often to identify which policy – out of a range of possible design or implementation options – will have the greatest impact on the outcome in question. RCTs require painstaking and costly data collection with potentially long periods of observation.

Kasy and Sautmann, (2020) use an Adaptive experiment design combining 'multi-armed bandits' from the reinforcement learning literature and propose an explorative sampling design.

Combining Econometrics with Machine Learning

Precision Agriculture for Development (PAD) - an NGO, provides a free agricultural extension service for smallholder farmers. For an enrollment drive in India with one million farmers, PAD wanted to learn as quickly as possible how to conduct enrollment calls effectively, so that farmers would not screen the calls without learning about the service.

Combining Econometrics with Machine Learning

Precision Agriculture for Development (PAD) - an NGO, provides a free agricultural extension service for smallholder farmers. For an enrollment drive in India with one million farmers, PAD wanted to learn as quickly as possible how to conduct enrollment calls effectively, so that farmers would not screen the calls without learning about the service.

Using a reinforcement learning approach PAD identified calling at 10 am with a text message one hour ahead emerged as the most successful treatment early on.

Conclusion

ML in economics is more than the sum of its parts. It is much more than mere Applied ML.

Conclusion

ML in economics is more than the sum of its parts. It is much more than mere Applied ML.

Econometrics is the study of measuring economic concepts. The field has development numerous techniques over the last 50 years that proved to be immensely useful. They are used by other fields as well e.g. epidemiology, neuroscience, sociology, business, finance, political science.

Conclusion

ML in economics is more than the sum of its parts. It is much more than mere Applied ML.

Econometrics is the study of measuring economic concepts. The field has development numerous techniques over the last 50 years that proved to be immensely useful. They are used by other fields as well e.g. epidemiology, neuroscience, sociology, business, finance, political science.

The combination of ML and Econometrics is just the begining of many more exciting tools and contributions!