Machine Learning for the Social Sciences

Comparing ML and traditional econometric approaches

Andrew Peterson

10.04.2018

Université de Genève

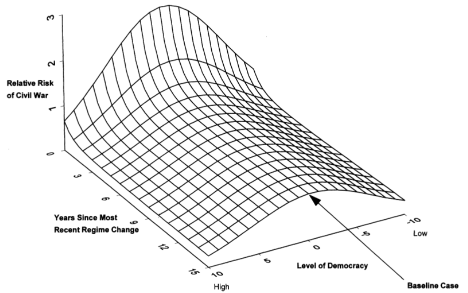
Introduction

Table of contents

Motivation

- Assumptions are not just technical details – they may drive conclusions & research agendas, methodologies
- e.g. do we need more data? What kind?
- e.g. more murder in the middle?

FIGURE 1. Relative Risk of Civil War as a Function of Democracy and Time since Most Recent Regime Change, 1816–1992



Note: The figure is based on the parameter estimates in Table 2B. The baseline case is an observation with democracy = 0 and proximity of regime change = 0 (15 years since regime change). All risks are plotted relative to this case. For instance, an observation with democracy = 0 and years since most recent regime change = 0 is estimated to have a risk of exp($\hat{\beta}$, exp(−0/527) + 0$\hat{\beta}_2$) = exp($\hat{\beta}_1$) = 2.68 relative to the baseline (cf. note 12).

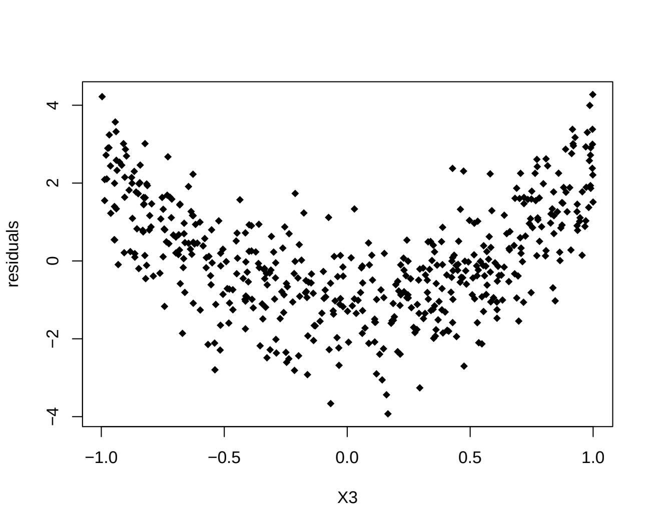
Source: Hegre, et al. 2001

A very simple example

- Say the true DGP involved a squared term but we failed to include it in our model.
- True model: $y = x_1 + x_2 + x_3^2 + \epsilon$
- We estimate OLS: $y = x_1 + x_2 + x_3 + \epsilon$
- What will happen? How might we know if there is a problem?

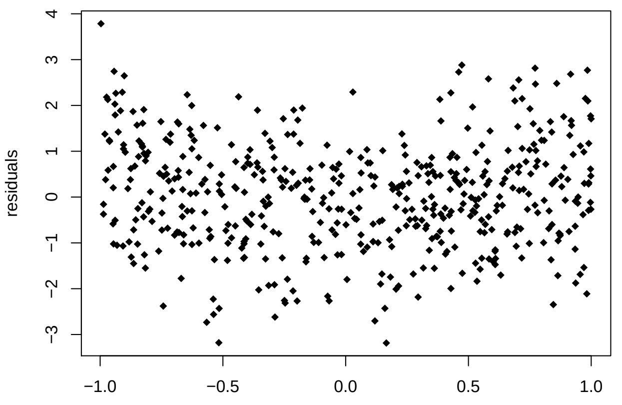
Simple example

› If we're careful, we would look at the residuals (plotted against x_3 and see there is a problem:



Simple example

- By contrast, if we just threw the data in, say, a random forest model and generated predictions, model takes care of it (MSE 1.34 vs 2.12, (though problems with interpreting coefficients...):



ML vs ’metrics: A rough first take* I

Advantages of machine learning

- “data-driven model selection”: Classical view assumes the researcher has the true functional form. Since they generally do not, in practice they may consider models on the basis of other criteria, e.g. p-hacking.
- Failure to get the model right may lead to inconsistent or false interpretations (as when a variable has a non-linear effect, e.g. if education is included as years of education but PhDs earn less than MAs...)
- Improvements in predictive accuracy, ability to develop new measures.

  * (We will see some of these views evolve below.)

ML vs 'metrics: A first take I

But:

- Econometrics has solutions to the first two of these. e.g. dummy-ing out variables, multiple comparison corrections, conditions under which linearity assumption is critical ... (more below)

ML vs 'metrics: A first take II

Disadvantages of machine learning

- β coefficients are unstable, difficult to interpret
- OVB-like bias in other estimators (Mullainathan &
- no standard errors are provided
- big data, ML – but that does nothing to address the causal identification problem

ML vs 'metrics: A first take II

But:

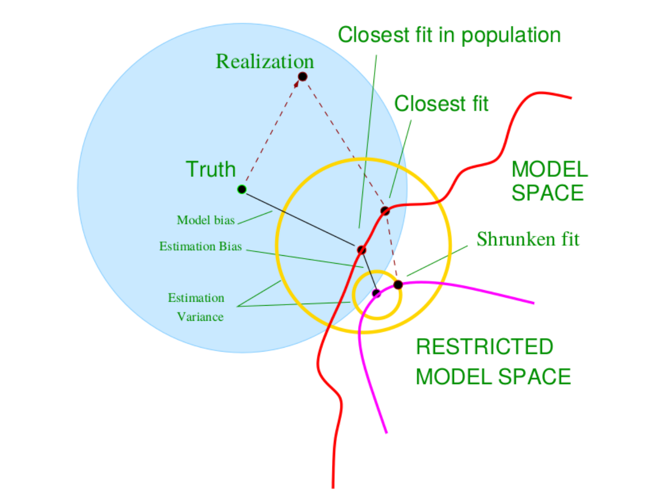
- Can address robustness of coefficients
- Can develop standard errors
- ML can act in the service of a causal inference

ML: Assumptions & Approaches

ML Bias and Variance

- Bias induced by functional form (model space)
- Variance introduced in estimating the model
- Bias induced by regularization

ML Bias and Variance



Source: Hastie et al. ESL

ML assumptions

- iid draws of y, X (so training data reflects test data)
- restrictions on function class
- guarantees about optimal regularization through cross-validation, etc
- optimization process achieves a global/good minimum

Assumptions: Regression inference

What assumptions being made in machine learning & econometric models?

How do (or *should*) these translate into methodological approaches?

Linear Regression

- First, quick review of assumptions underlying regression
- Then we will consider the causal identification part

Regression as Parametric Model

- $Y_i = X_i'\beta + \epsilon$ (linear relationship, X is full rank)

  where Y_i is iid (conditionally independent...)

- "spherical errors": $\epsilon \sim$ N(0,σ^2)

- Solution: $\hat{\beta} = (X'X)^{-1}X'y$,

  standard errors from $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}$

Regression as Semi-Parametric Model

- Conditional moment restrictions:

$$\mathbb{E}[\epsilon|X] = 0 \tag{1}$$

$$\text{cov}[\epsilon|X] = \sigma^2 \cdot \mathbb{I}_{nxn} \tag{2}$$

- OLS is BLUE
- or adapt for heteroskedasticity, limited DV, etc...

Agnostic Regression & Approximation inference

- Classical regression inference: assume correct specification !
- Agnostic regression: We can get a consistent estimate of the linear approximation, even if CEF is not perfect
- (i.e. with more data, converges in probability to the true value, but think about what this means when the relationship is nonlinear...)

Conditional expectation function

- Conditional expectation function

  e.g. how does pay vary with education?
- CEF is multivariate function *m*:

  $\mathbb{E}(wage|sex, race, education) = m(x_1, x_2, x_3)$

Agnostic Regression

- "...the linear CEF model is empirically unlikely to be accurate unless x is discrete and low-dimensional so all interactions are included. Consequently in most cases it is more realistic to view the linear specification as an approximation."
  (Hanson, 2018. Econometrics)

Agnostic Regression: Example

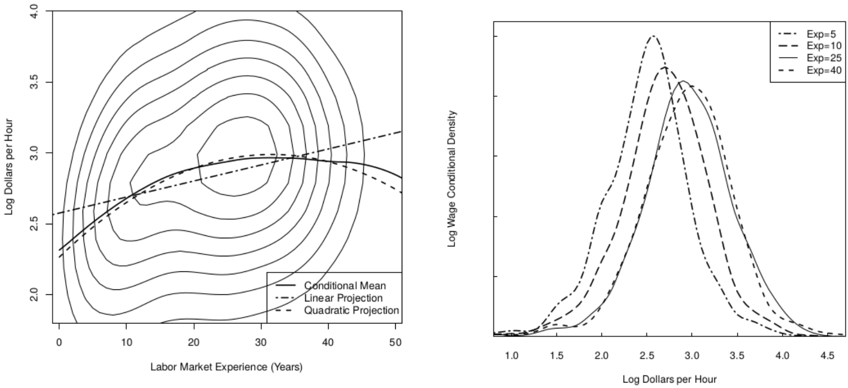
- Returns to additional experience? (Hanson 2018)
- linear projection

$$Proj(log(wage)|experience) = 2.5 + 0.011 Experience$$

- quadratic projection

$$Proj(log(wage)|experience) = 2.3 + 0.046 Experience - 0.0007 Experience^2$$

Agnostic Regression: Example



(a) Joint density of log(wage) and experience and conditional mean

(b) Conditional density

Source: Hanson, 2018. Econometrics.

Causality and Regressions

When can a regression estimate be used to estimate a causal effect?

Potential outcomes and causality

- NB: As you may not be familiar with the potential outcomes framework, in what follows I have avoided using the formal notation.
- In any case, my purpose here is simply to get us thinking more carefully about what we hope to get out of regression analysis.
- That said, a more formal treatment would allow more precision and insight (if people are interested, this could be the topic of another session, otherwise see the full sources I drew on, in particular Angrist & Pischke MHE (here), Cyrus Samii slides (here), or Hanson, Econometrics textbook (here)).

Potential outcomes and causality

- RQ: How does a treatment (e.g. seeing an ad) change the probability of turning out to vote (for the average person, or females of a given age, etc).
- Problems: (1) treatment effect varies by individual, (2) for each individual, we only observe one outcome (e.g. whether Bob voted after seeing the message, but not whether he voted after *not* seeing the message.
- Say we estimate the regression:

$$voted_i = \beta_1 ad_i + \mathbf{X}_i \beta_2 + \epsilon_i$$

- When can we interpret the regression coefficient (β_1) as a causal effect? (for whom and under what circumstances)?

Identification

How could two *identical* individuals be equally likely to receive the treatment?

- random assignment
- unconfoundedness (my focus here, e.g. bureaucratic flukes, ...)

Potential outcomes and causality

"A regression is causal when the CEF it approximates is causal.

The CEF is causal when it describes differences in average potential outcomes for a fixed reference population." (Angrist & Pischke, MHE)

Unconfoundedness / Conditional Independence Assumption

- treatment assignment is independent of potential outcomes conditional on X
- e.g. conditional on parents' income & other X, the likelihood of having a high-school diploma is independent of intelligence and other unobserved vars that affect income

Potential outcomes and causality

- random draw from population, SUTVA
- unconfoundedness / CIA
- linearity in parameters (X_i)
- constant effects (treatment effect same for all individuals)

Regression estimates of causal effects

- If all of these are met, regression can estimate causal effects
- If heterogeneous effects, the regression estimate is biased and inconsistent (variance-weighted average), but
- under some conditions this can be corrected by regession weighting (Samii & Aronow 2016 AJPS)
- When confounding not linear in X_i, additional bias problem.

So what?

- Some issues (e.g. identification) are orthogonal to ML
- Model specification issues raised about linear regression can be exacerbated or aided by ML, as we will see next...

Problems with $\hat{\beta}$ estimates (Mullainathan & Spiess 2017)

- instability (regularization)
- OVB-like bias in other estimators (regularization)
- lack of s.e.

We will consider these in Part III.

3 categories of applications (Mullainathan & Spiess 2017)

- use $\hat{y}$
- use $\hat{y}$ to better predict $\hat{\beta}$ (ML for Causal inference)
- direct policy applications

We will come back to these in Part IV.

Questions?