Machine Learning for Economics and Social Sciences: Applications and Software

Achim Ahrens (ETH Zürich)
achimahrens.de
achim.ahrens@gess.ethz.ch

Applied Machine Learning Days, March 30, 2022

Introduction

About me

- Senior Data Scientist & Post-doctoral Researcher at the Public Policy Group, ETH Zurich and Immigration Policy Lab (ETH/Stanford)
- Background in economics & econometrics
- Tasks: Data science support and research
- "Hobby:" Develop statistical software packages

Introduction

About me

- Senior Data Scientist & Post-doctoral Researcher at the Public Policy Group, ETH Zurich and Immigration Policy Lab (ETH/Stanford)
- Background in economics & econometrics
- Tasks: Data science support and research
- "Hobby:" Develop statistical software packages

Public Policy Group / Immigration Policy Lab

- Policy evaluation with a focus on immigration topics
- Combining survey & registry data with causal inference & ML
- Example projects:
  - Counterspeech strategies for hate speech (Hangartner et al., 2021)
  - Discrimination on online recruitment platforms (Hangartner, Kopp, and Siegenthaler, 2021)
  - Effect of citizenship on immigrants (Hainmueller, Hangartner, and Ward, 2019)

Today's talk

I'll talk about:

- The (increasing) importance of ML in economics & social sciences
- Combining ML & causal inference
- Challenges of writing accessible software (for non-ML experts)

Empirical economics and social sciences

- In recent years, machine learning (ML) has increasingly been leveraged in social sciences and economics.
- Sometimes, ML tools can be used "off the shelf" (e.g. predicting long-term unemployment; Mullainathan and Spiess, 2017),...

Empirical economics and social sciences

- In recent years, machine learning (ML) has increasingly been leveraged in social sciences and economics.
- Sometimes, ML tools can be used "off the shelf" (e.g. predicting long-term unemployment; Mullainathan and Spiess, 2017),...
- but most often research question is of *causal nature*.
- *Example:*

  Say you want to predict hotel occupancy rates. High price predicts high hotel occupancy, but that doesn't mean hoteliers should increase prices to increase occupancy (Athey, 2017).

Empirical economics and social sciences

- *Typical research question:* What's the effect of policy X on outcome Y?
  - The effect of a refugee policies on long-term labour market integration
  - The effect of citizenship on wages/employment

Empirical economics and social sciences

- *Typical research question:* What's the effect of policy X on outcome Y?
  - The effect of a refugee policies on long-term labour market integration
  - The effect of citizenship on wages/employment
- *Methods:* Quasi-experiments, Difference-in-differences, Regression Discontinuity, Instrumental Variables
  - focus on *identification strategies*: research designs that yield causal treatment effects
  - "Credibility revolution" (Angrist and Pischke, 2010; Nobel Prize in Economics 2021)

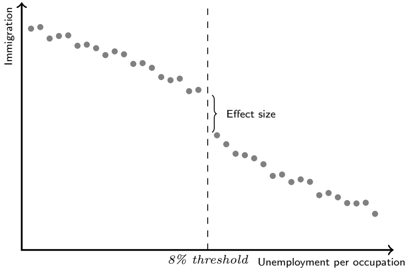
Application: Evaluation of Swiss prioritisation policy

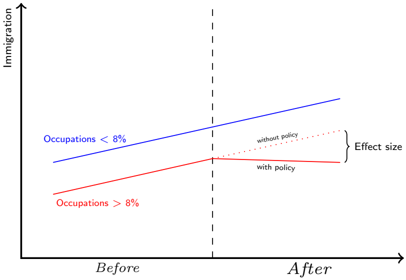
Policy:

- In February 2014 the initiative 'Against mass immigration' was adopted by Swiss voters.
- Switzerland introduced a set of policies affecting occupations with unemployment rate above 8%. Policies aimed at improving hiring of domestic work force relative to outside workers.

Concerns: Labor market outcomes and immigration is confounded by seasonal and business-cycle effects. A simple before-after comparison is not enough.

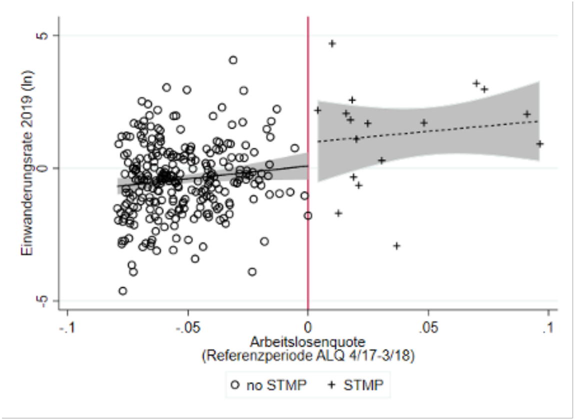
Application: Evaluation of Swiss prioritisation policy





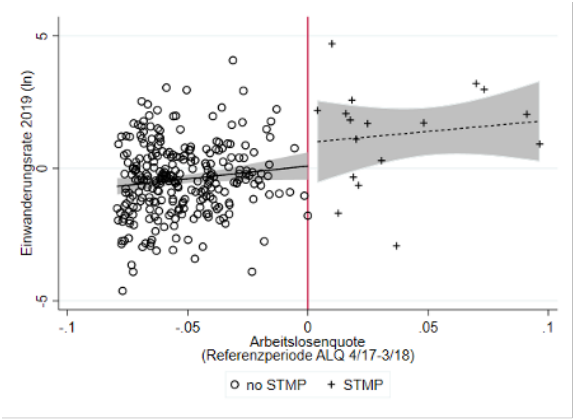
Methodology: We combine *Regression Discontinuity Design* (left) and *Difference-in-diffences* (right).

Application: Evaluation of Swiss prioritisation policy



Identification strategy: The RD design exploits the quasi-random threshold at 8% and compares occupations around that threshold.

Application: Evaluation of Swiss prioritisation policy



Main results: Policy objective was overall not achieved. No effect unemployment and immigration (see Ahrens et al., 2021).

Supervised ML

- Generally focused on prediction/classification tasks rather than causal inference
- *Large toolbox:* regularized regression, random forest, SVM, neural nets, etc.
- *Procedure:* Algorithm is trained on some data and validated using unseen data.
- *Strengths:* *Out-of-sample* prediction/classification, high-dimensional data, data-driven model selection.

Causal ML

How ML can be used for *causal inference*?

Causal ML

How ML can be used for *causal inference*? — Three common approaches:

1. Exploring *treatment effect heterogeneity* (e.g. Causal Forests, GATES): which groups are most and least affected by a policy?
2. *Policy learning:* Who should become which treatment?
3. *Robust causal inference* in the presence of high-dimensional controls and/or instruments

Causal ML

How ML can be used for *causal inference*? — Three common approaches:

1. Exploring *treatment effect heterogeneity* (e.g. Causal Forests, GATES): which groups are most and least affected by a policy?
2. *Policy learning:* Who should become which treatment?
3. *Robust causal inference* in the presence of high-dimensional controls and/or instruments

I will focus on 3 today.

Causal ML

Motivating example. The partial linear model:

$$y_i = \underbrace{\boxed{\theta d_i}}_{\text{causal part}} + \underbrace{\boxed{g(\boldsymbol{x}_i)}}_{\text{nuisance}} + \varepsilon_i.$$

where

- d_i is a treatment or policy variable.
- We want to estimate the causal effect θ.
- However, we causal effect is only plausible if we control for observed variables $\boldsymbol{x}_i$.

Causal ML

Motivating example. The partial linear model:

$$y_i = \underbrace{\boxed{\theta d_i}}_{\text{causal part}} + \underbrace{\boxed{g(\boldsymbol{x}_i)}}_{\text{nuisance}} + \varepsilon_i.$$

Application:

- Hangartner, Kopp, and Siegenthaler (2021) assess hiring discrimination on job platform
- y_i = contact rate; d_i = minority indicator; $\boldsymbol{x}_i$ = job seeker characteristics
- We can only interpret the discrimination effect θ as causal once we have controlled for all job seeker characteristics that are observable to the recruiter on the job platform ($\boldsymbol{x}_i$).

Causal ML

Motivating example. The partial linear model:

$$y_i = \underbrace{\boxed{\theta d_i}}_{\text{causal part}} + \underbrace{\boxed{g(\boldsymbol{x}_i)}}_{\text{nuisance}} + \varepsilon_i.$$

How do we account for confounding factors $\boldsymbol{x}_i$*?* — The standard approach is to assume linearity $g(\boldsymbol{x}_i) = \boldsymbol{x}_i'\beta$ and employ ordinary least squares.

Causal ML

Motivating example. The partial linear model:

$$y_i = \underbrace{\theta d_i}_{\text{causal part}} + \underbrace{g(\boldsymbol{x}_i)}_{\text{nuisance}} + \varepsilon_i.$$

How do we account for confounding factors $\boldsymbol{x}_i$? — The standard approach is to assume linearity $g(\boldsymbol{x}_i) = \boldsymbol{x}_i'\beta$ and employ ordinary least squares.

Problems:

- Non-linearity & unknown interaction effects
- High-dimensionality: we might have "many" controls
- We don't know which controls to include

Double ML

Motivating example. The partial linear model:

$$y_i = \underbrace{\theta d_i}_{\text{causal part}} + \underbrace{g(\boldsymbol{x}_i)}_{\text{nuisance}} + \varepsilon_i.$$

One solution: Double-ML employs two auxiliary estimations of $y_i \rightsquigarrow \boldsymbol{x}_i$ and $d_i \rightsquigarrow \boldsymbol{x}_i$ to extract the effect of $\boldsymbol{x}_i$ on y_i.

Two flavours:

- Use Lasso with 'theoretically justified' penalization

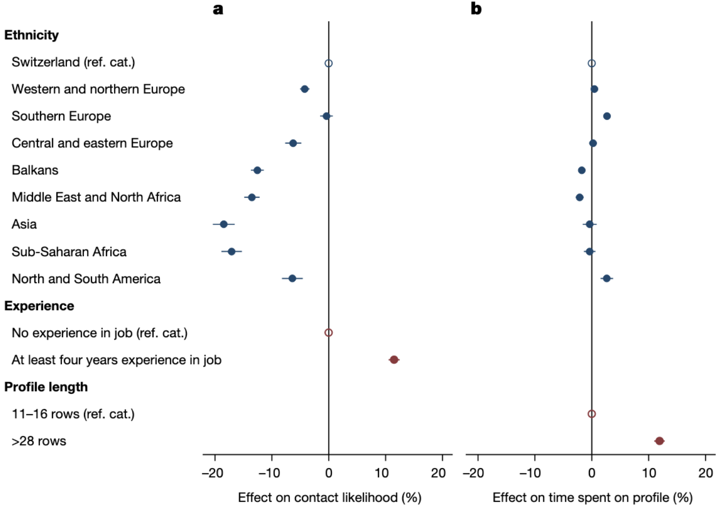
  Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov, Hansen, and Spindler, 2015

- Use 'any' machine learner with sample-splitting

  Chernozhukov et al., 2018

Double ML approaches are quite general: multiple treatment variables, IV settings, LATE estimation.

Double ML



Main finding: contact rates are lower by 4-19% lower for individuals from immigrant and minority ethnic groups (Hangartner, Kopp, and Siegenthaler, 2021).

Adoption

Double ML approaches are now increasingly used as a central identification strategy, complementary robustness check or simply to gain precision in RCTs.

- Covid transmission (Qiu, Chen, and Shi, 2020), health outcomes (Jones, Molitor, and Reif, 2019), development programs (Hess, Jaimovich, and Schündeln, 2020)

Adoption of ML & Double ML is partially hindered by lack of knowledge, but also by *availability of software*:

$\rightarrow$ Stata is along with R most widely used software in economics and social science, but has only limited ML features

Software

We have implemented Double-Lasso approaches for Stata (Ahrens, Schaffer, Hansen, 2018, 2019, 2022):

- `lassopack` for regularized regression
- `pdslasso` for Double-Lasso approaches
- `ddml` for Double-ML with sample splitting
- `pystacked`: scikit-learn front-end with a focus on stacking

Software

We have implemented Double-Lasso approaches for Stata (Ahrens, Schaffer, Hansen, 2018, 2019, 2022):

- `lassopack` for regularized regression
- `pdslasso` for Double-Lasso approaches
- `ddml` for Double-ML with sample splitting
- `pystacked`: scikit-learn front-end with a focus on stacking

Similar implementations exists for R & Python (Chernozhukov, Hansen, and Spindler, 2016; Bach et al., 2021, Microsoft's `EconML`).

Considerations

Contact with users has also highlighted risks:

- Focus on single learner (often Lasso)
- No or insufficient tuning
- Defaults have a huge effect on user behaviour
  $\rightarrow$ Defaults need to be well chosen and justified

Question: What would be a sensible default ML method?

Stacking regression

Which machine learner should we use?

We suggest *Stacking regression* (Wolpert, 1992; Breiman, 1996) as the *default* machine learner, which we have implemented in the separate program `pystacked` using Python's scikit learn.

Stacking is an ensemble method that combines multiple base learners into one model. As the default, we use *non-negative least squares*:

$$\boldsymbol{w} = \arg\min_{w_j \geq 0} \sum_{i=1}^{n} \left(y_i - \sum_{m=1}^{M} w_m \hat{f}_m^{-i}(\boldsymbol{x}_i) \right)^2 ,$$

where $\hat{f}_m^{-i}(\boldsymbol{x}_i)$ are cross-validated predictions of base learner m.

Does it make a difference?

Let's do a simple simulation experiment.

We generate artificial data according to two data-generating processes:

$$y_i = \theta d_i + g(\boldsymbol{x}_i) + \varepsilon_i.$$

where

1. $g(\boldsymbol{x}_i)$ is linear
2. $g(\boldsymbol{x}_i)$ is generated to be non-linear

Aim: we want to estimate the causal effect θ.

Double-ML + Stacking

	$n_s = 9915$			$n_s = 99150$		
Panel (A): Linear DGP	Bias	MAB	Rate	Bias	MAB	Rate
Full sample:						
OLS	100.99	918.03	.95	-22.61	255.52	.94
PDS-Lasso	101.83	913.18	.95	-19.9	257.29	.94
DDML methods:						
Base learners						
OLS	105.07	906.96	.94	-23.05	256.51	.94
Lasso with CV (2nd order poly)	104.33	907.84	.94	-22.45	257.23	.94
Ridge with CV (2nd order poly)	103.22	898.56	.94	-23.27	255.54	.94
Lasso with CV (10th order poly)	49.56	1120.59	.93	37.98	260.53	.95
Ridge with CV (10th order poly)	1066	1342.38	.9	15.85	260.41	.95
Random forest (low regularization)	-59.63	1083.64	.91	-59.29	343.46	.86
Random forest (high regularization)	105.58	952.35	.94	-46.54	275.56	.91
Gradient boosting (low regularization)	53.97	930.93	.94	-41.84	252.14	.94
Gradient boosting (high regularization)	162.75	923.08	.95	48.31	259.12	.95
Meta learners						
Stacking: NNLS	100.01	935.27	.94	-22.7	254.01	.94

DDML with Stacking does equally well as OLS when the DGP is linear. . .

Double-ML + Stacking

	$n_s = 9915$			$n_s = 99150$		
Panel (A): Linear DGP	Bias	MAB	Rate	Bias	MAB	Rate
Full sample:						
OLS	-2496.16	2477.19	.63	-2658.04	2636.31	0
PDS-Lasso	-2507.47	2489.77	.62	-2657.5	2635.94	0
DDML methods:						
Base learners						
OLS	-2522.98	2540.36	.62	-2660.54	2640.98	0
Lasso with CV (2nd order poly)	767.2	1078.29	.91	691.67	695.3	.64
Ridge with CV (2nd order poly)	825.21	1091.19	.9	702.55	707.28	.64
Lasso with CV (10th order poly)	-4214.09	1895.22	.92	-10.06	294.34	.94
Ridge with CV (10th order poly)	-2123.59	2095.56	.91	4.42	288.37	.94
Random forest (low regularization)	-104.54	1019.55	.92	-28.83	332.87	.87
Random forest (high regularization)	-110.06	959.96	.95	-21.52	280.36	.94
Gradient boosting (low regularization)	69.44	890.94	.95	7.28	263.62	.95
Gradient boosting (high regularization)	213.04	895.47	.95	174.14	291.63	.93
Meta learners						
Stacking: NNLS	-62.97	1068.87	.84	18.36	269.02	.95

…but also yields lowest bias when the DGP is non-linear.

Summary

Machine Learning is moving into the *standard toolbox* in economics & social sciences.

Yet, given that research questions are usually of *causal nature*, ML can often not be applied 'off the shelf.'

Double ML approaches are an example of a synthesis between ML and causal inference. Other examples: Heterogeneous treatment effects, policy learning.

Thanks for listening

Contact:

`achim.ahrens@gess.ethz.ch`

References I

- Ahrens, Achim, Christian B Hansen, and Mark E Schaffer (Feb. 2018). *PDSLASSO: Stata Module for Post-Selection and Post-Regularization OLS or IV Estimation and Inference*. URL: https://ideas.repec.org/c/boc/bocode/s458459.html.
- Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer (Jan. 16, 2019). *Lassopack: Model Selection and Prediction with Regularized Regression in Stata*. URL: http://arxiv.org/abs/1901.05397 (visited on 01/17/2019).
- Angrist, Joshua D. and JÃ¶rn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics". In: *Journal of Economic Perspectives* 24.2, pp. 3–30. URL: https://www.aeaweb.org/articles?id=10.1257/jep.24.2.3.
- Athey, Susan (2017). "Beyond prediction: Using big data for policy problems". In: *Science* 355.6324, pp. 483–485.

References II

- Bach, P. et al. (2021). *DoubleML – An Object-Oriented Implementation of Double Machine Learning in R*. arXiv:2103.09603 [stat.ML].
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on Treatment Effects after Selection among High-Dimensional Controls". In: *Review of Economic Studies* 81, pp. 608–650. URL: https://doi.org/10.1093/restud/rdt044.
- Breiman, Leo (Aug. 1996). "Bagging Predictors". In: *Machine Learning* 24.2, pp. 123–140. URL: https://doi.org/10.1007/BF00058655.
- Chernozhukov, Victor, Chris Hansen, and Martin Spindler (2016). "hdm: High-Dimensional Metrics". In: *The R Journal* 8.2, pp. 185–199. URL: https://doi.org/10.32614/RJ-2016-040.

References III

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (May 2015). "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments". In: *American Economic Review* 105.5, pp. 486–490. URL: https://doi.org/10.1257/aer.p20151022.

Chernozhukov, Victor et al. (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097.

Hainmueller, Jens, Dominik Hangartner, and Dalston Ward (2019). "The effect of citizenship on the long-term earnings of marginalized immigrants: Quasi-experimental evidence from Switzerland". In: *Science Advances* 5.12, eaay1610. URL: https://www.science.org/doi/abs/10.1126/sciadv.aay1610.

References IV

- Hangartner, Dominik, Daniel Kopp, and Michael Siegenthaler (Jan. 2021). "Monitoring hiring discrimination through online recruitment platforms". In: *Nature* 589.7843, pp. 572–576. URL: https://doi.org/10.1038/s41586-020-03136-0.
- Hangartner, Dominik et al. (2021). "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment". In: *Proceedings of the National Academy of Sciences* 118.50, e2116310118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2116310118.
- Hess, Simon, Dany Jaimovich, and Matthias Schündeln (June 2020). "Development projects and economic networks: Lessons from rural gambia". In: *The Review of Economic Studies* 88.3. tex.eprint: https://academic.oup.com/restud/article-pdf/88/3/1347/38107873/rdaa033.pdf, pp. 1347–1384. URL: https://doi.org/10.1093/restud/rdaa033.

References V

Jones, Damon, David Molitor, and Julian Reif (2019). "What do workplace wellness programs do? Evidence from the Illinois workplace wellness study". In: *The Quarterly Journal of Economics* 134.4, pp. 1747–1791.

Mullainathan, Sendhil and Jann Spiess (May 2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106. URL: http://www.aeaweb.org/articles?id=10.1257/jep.31.2.87.

Qiu, Yun, Xi Chen, and Wei Shi (2020). "Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China". In: *Journal of Population Economics* 33.4, pp. 1127–1172.

Wolpert, David H. (1992). "Stacked Generalization". In: *Neural Networks* 5.2, pp. 241–259. URL: https://www.sciencedirect.com/science/article/pii/S0893608005800231.