Introduction to Causal Machine Learning

Anthony Strittmatter

Lecturer

Anthony Strittmatter

Research Interests: Business, Labour, and Health Economics, Program Evaluation, Computational Data Analytics

Positions:

Since 2023	Full Professor for Applied Econometrics at UniDistance Suisse in Brig/Valais
2022-2024	Senior Economist at Amazon in London
2020-2023	Institut Polytechnique/CREST in Paris
2014-2020	University of St.Gallen, with research visits at UC Berkeley, Stanford University, and LMU Munich
2009-2016	Albert-Ludwig University of Freiburg

- Email: anthony.strittmatter@unidistance.ch
- Webpage: www.anthonystrittmatter.com

Causal ML Lectures

- **École Polytechnique Paris**, France, 2024
- **University of Basel**, Switzerland 2024, 2022, 2021, 2020
- **ENSAE Paris**, France, 2024, 2023, 2022, 2021
- **University of Lucerne**, Switzerland, 2022, 2021
- **Johannesburg University**, South Africa, 2022
- **Central Bank of Estonia**, Tallin, 2021
- **University of Würzburg**, Germany, 2021
- **University of Duisburg-Essen**, Germany, 2021
- **CESifo Munich**, Germany, 2019
- **University of Hohenheim**, Germany, 2019
- **ZEW Mannheim**, Germany, 2020

Schedule

	Lecture 1	**Lecture 2**	**Lecture 3**	**Lecture 4**
Date	**Monday August 19**	**Tuesday August 20**	**Wednesday August 21**	**Thursday August 22**
Time	10:15-12:00 13:15-15:00	10:15-12:00 13:15-15:00	10:15-12:00 13:15-15:00	10:15-12:00 13:15-15:00
Room	EC3:108	EC3:108	EC3:108	EC3:108

Course Outline

- **Lecture 1:** Introduction to Statistical Learning
  - Prediction vs. Causal Analysis
  - Regularized Regression: Lasso, Ridge, Elastic Net
- **Lecture 2:** Non-parametric Supervised Machine Learning
  - Trees and Random Forests
  - Deep Learning
- **Lecture 3:** High-Dimensional Confounding
  - Double Selection Procedure
  - Double Machine Learning
- **Lecture 4:** Effect Heterogeneity and Policy Learning
  - Causal Forest
  - Optimal Policy Learning
  - Bandit Algorithms

PC Labs

- PC labs are an integral component of this course, providing hands-on experience with machine learning tools and techniques.
- All course materials are available on JupyterHub, accessible via your personalized link: http://54.196.158.125/user/<email address>
- Notebooks eliminate the need for software installation and ensure data is correctly organized for each session.
- For your own research projects, we recommend installing R and RStudio on your local PC.

Grading: Research Proposal

- Individual Home Assignment
- Deadline for submission: September 29, 2024
- Grades: Pass/Fail

General

- Feel free to interrupt me at any time with questions.
- Let me know if I'm too slow or too fast. Ask me to repeat material if something was not clear.
- You can also send me an email with questions: anthony.strittmatter@unidistance.ch
- Suggestions for course improvement are welcome.
- Please interact with your fellow students and build learning groups.

References

- Mullainathan and Spiess (2017): "Machine Learning: An Applied Econometric Approach", Journal of Economic Perspectives, 31 (2), pp. 87-106, download.
- Athey (2017): "Beyond Prediction: Using Big Data for Policy Problems", Science, 355 (6324), pp. 483-485, download.

What is Machine Learning (ML)?

- ML (or statistical learning) methods have existed for decades.
- Currently, "Machine Learningïs a buzzword with no clear definition.
- Most people think of ML as computationally intensive methods that make data-driven modeling decisions and/or can deal with large data amounts.
- However, relevant textbooks consider even OLS/Logit as statistical learning tools.

Purpose of Machine Learning

- Consider the structural model

$$Y = f(X) + \epsilon = X\beta + \epsilon,$$

  with $E[\epsilon] = 0$.
- Causal analysis aims to estimate $\hat{\beta}$, with $plim(\hat{\beta}) = \beta$.
- **Machine learning** aims to predict Y.
- There is a clear link between causal analysis and machine learning, because

$$\widehat{Y} = \hat{f}(X) = X\hat{\beta}$$

  is a potential predictor for Y.
- Parameter consistency does not have the highest priority when it comes to predictions.

Potential Advantages and Disadvantages of ML

- ML methods can be very powerful in predicting Y, even when $\hat{\beta}$ is biased.
- ML methods can incorporate many (or even high-dimensional) covariates X conveniently.
- ML methods can model $\hat{f}(\cdot)$ flexibly and data-driven.
- **Main disadvantage:** ML is a black-box approach, and we lose the interpretability of $\hat{f}(\cdot)$ or $\hat{\beta}$.

Prediction vs. Causality

Causal vs. Predictive Questions

Predictive Questions:

- How will the oil price change tomorrow (forecasting)?
- How high is the current unemployment rate (nowcasting)?
- Which adolescents have a high probability of becoming addicted to drugs (policy prediction)?

Causal Questions:

- What is the effect of a tweet by President Donald Trump on oil prices?
- How does inflation affect the unemployment rate?
- Can prevention programs reduce the probability of drug addiction among high-risk youths?

Assessing the Model Accuracy

Causal Analysis:

- True β is unobservable.
- Assess the model with asymptotic properties

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2).$$

- Finite sample biases are mostly neglected.

Assessing the Model Accuracy

Prediction:

- We observe Y for each unit (e.g., individual).
- We can assess the model accuracy directly in the sample of our analysis, for example, using the mean-squared-error (MSE)

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2.$$

- MSE accounts for finite sample biases.

Example: Prediction of Used Car Prices

- We have access to web-scraped data from the online advertisement platform *myLemons*.
- We want to predict asking prices of used cars based on observable characteristics.
- We observe around 40 covariates about car brand, mileage, age, emissions, maintenance certificate, seller type, guarantee, etc. (including several non-linear and interaction terms)

In-Sample MSE

- Partition data into training and test samples
- In the training sample, we estimate the empirical model

$$Y_{tr} = \hat{f}_{tr}(X_{tr}) + \hat{\epsilon}_{tr} = X_{tr}\hat{\beta}_{tr} + \hat{\epsilon}_{tr}$$

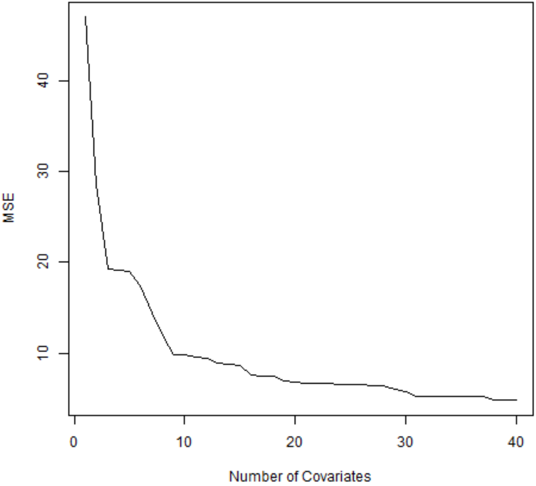
- In the training sample, we predict the fitted values

$$\widehat{Y}_{tr} = \hat{f}_{tr}(X_{tr}) = X_{tr}\hat{\beta}_{tr}$$

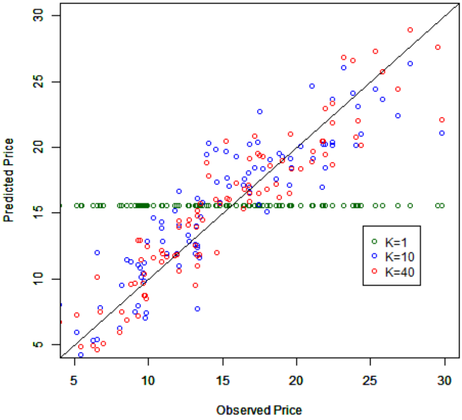
and calculate the MSE

$$\widehat{MSE}_{tr} = \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} (Y_{i,tr} - \widehat{Y}_{i,tr})^2.$$

MSE in Training Sample



Predicted Car Prices in Training Sample



Number of Covariates	1	10	40
MSE	46.948	9.819	4.866

Out-of-Sample MSE

- In the training sample, we estimate the empirical model

$$Y_{tr} = \hat{f}_{tr}(X_{tr}) + \hat{\epsilon}_{tr} = X_{tr}\hat{\beta}_{tr} + \hat{\epsilon}_{tr}$$

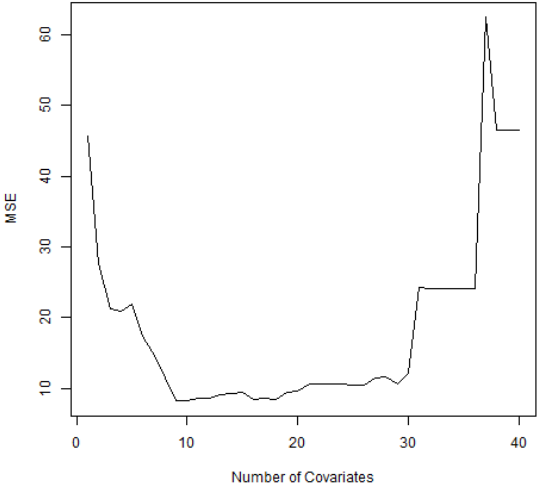
- In the test sample, we predict the fitted values

$$\widehat{Y}_{te} = \hat{f}_{tr}(X_{te}) = X_{te}\hat{\beta}_{tr}$$

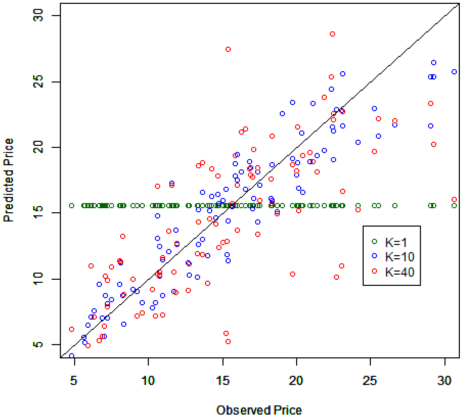
and calculate the MSE

$$\widehat{MSE}_{te} = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} (Y_{i,te} - \widehat{Y}_{i,te})^2.$$

MSE in Test Sample



Predicted Car Prices in Test Sample



Number of Covariates	1	10	40
MSE	45.742	8.222	46.499

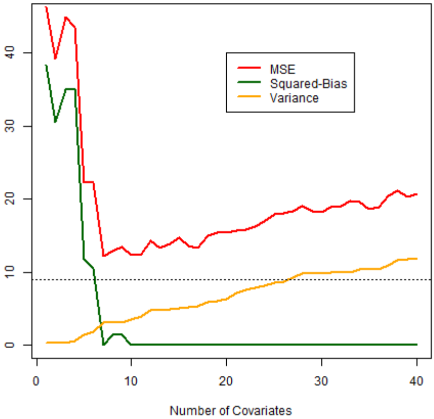
Bias-Variance Trade-Off

- When we assess the model for one randomly drawn individual from the test sample with fixed characteristics x_{te}, then we can decompose the MSE to

$$\begin{aligned}
MSE_{te} =& E[(Y_{te} - \widehat{Y}_{te})^2] \\
=& E[(f(x_{te}) + \epsilon_{te} - \hat{f}_{tr}(x_{te}))^2] \\
=& \underbrace{E[(f(x_{te}) - \hat{f}_{tr}(x_{te}))^2]}_{\text{Reducible}} + \underbrace{Var(\epsilon_{te})}_{\text{Irreducible}} \\
=& \underbrace{E[f(x_{te}) - \hat{f}_{tr}(x_{te})]^2}_{\text{Squared-Bias}} + \underbrace{Var(\hat{f}_{tr}(x_{te}))}_{\text{Variance}} + Var(\epsilon_{te})
\end{aligned}$$

- For i.i.d. data, $\hat{f}_{tr}(\cdot)$ and ϵ_{te} are independent of each other.

Simulation of Bias-Variance Trade-Off



- Only the first ten covariates have an impact on car prices in the simulation.
- Horizontal dashed line is the simulated noise $Var(\epsilon_{te})$.

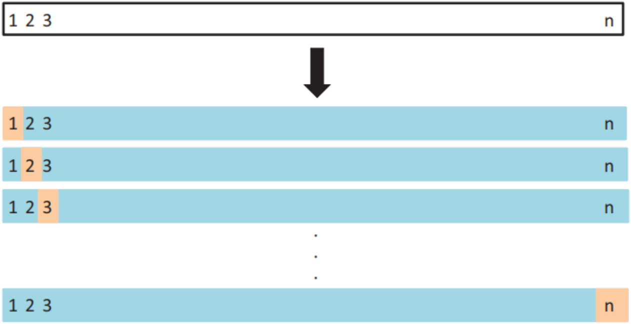
Lasso Example

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

	OLS	Lasso
Intercept	21.246	22.776
diesel	2.075	.
other_car_owner	0.730	.
pm_green	1.635	.
private_seller	6.100	0.076
guarantee	-2.440	-0.437
inspection	-0.813	.
maintenance_cert	1.481	.
mileage	-0.049	-0.031
age_car_years	-1.291	-1.012
R^2 training	0.655	0.543
R^2 test	0.606	0.611

Selection of Optimal Penalty Parameter

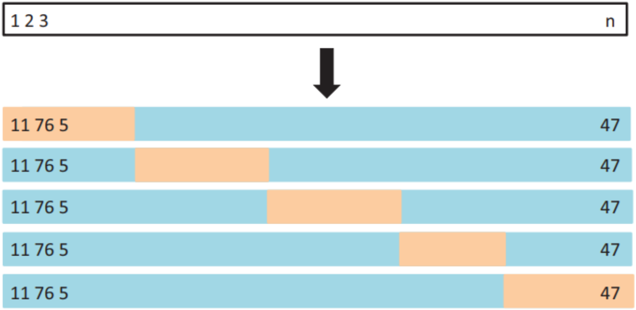
Leave-One-Out Cross-Validation



Source: James et al. (2013), p. 179

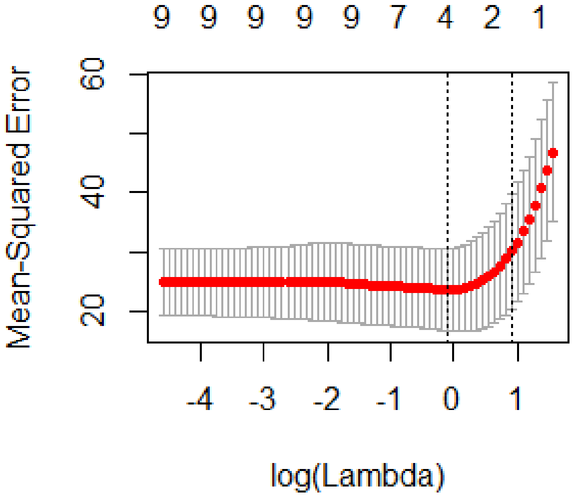
Selection of Optimal Penalty Parameter

k-fold Cross-Validation (CV) Algorithm



Source: James et al. (2013), p. 181

Cross-Validated MSE



Stability of the Lasso Model

	Lasso 1	Lasso 2	Lasso 3	Lasso 4	Lasso 5
Intercept	22.776	25.947	24.937	27.309	25.116
diesel	.	.	2.387	.	0.886
other_car_owner	.	-1.257	0.393	.	.
pm_green	.	2.871	.	.	.
private_seller	0.076	5.094	.	-1.037	.
guarantee	-0.437	1.677	15.939	.	.
inspection	.	-0.666	-0.374	.	.
maintenance_cert	.	-2.579	-0.868	.	.
mileage	-0.031	-0.037	-0.041	-0.069	-0.062
age_car_years	-1.012	-1.347	-1.416	-0.874	-1.115

$\rightarrow$ We do not learn the "trueßtructural model from ML

$\rightarrow$ ML is a black-box approach

Stability of the Lasso Predictions

Correlation of Predicted Car Prices in Test Sample:

	Lasso 1	Lasso 2	Lasso 3	Lasso 4
Lasso 2	0.94			
Lasso 3	0.85	0.81		
Lasso 4	0.97	0.91	0.85	
Lasso 5	0.99	0.94	0.87	0.99

When could Predictions be Useful?

Tasks with a prediction purpose:

- Predict stock or commodity prices using Twitter data.
- Nowcasting unemployment rate or GDP using Google search queries.
- Pre-screening of job applications.
- Consumer demand (shipping before the order occurs).
- Movie recommendations on Netflix.
- Handwriting, image, face, or voice recognition.

Examples of Business and Economic Studies

Prediction Tasks:

- Chandler, Levitt, and List (2011) predict shootings among high-risk youth to target mentoring interventions.
- Kleinberg, et al. (2018) predict the crime probability of defendants released from investigative custody to improve judge decisions.

Pre-Processing Unstructured Data:

- Glaeser et al. (2016) use images from Google Street View to measure block-level income in New York City and Boston.
- Kang et al. (2013) use restaurant reviews on Yelp.com to predict the outcome of hygiene inspections.
- Kogan et al. (2009) predict volatility of firms from market-risk disclosure texts (annual 10-K forms).

Predictions vs. Causal Inference

- Outcome (e.g., earnings): Y
- Binary Treatment (e.g., participation in training program): $D \in \{0, 1\}$
- Potential Outcome:
  - $Y(1)$ potential earnings under participation
  - $Y(0)$ potential earnings under non-participation

  $\rightarrow$ Only one potential earnings can be observed
- Causal effect: $\delta = Y(1) - Y(0)$

$\rightarrow$ Predictions have the observable estimation target $\widehat{Y}$

$\rightarrow$ Causal inference has the (partly) unobservable estimation target $\widehat{\delta}$

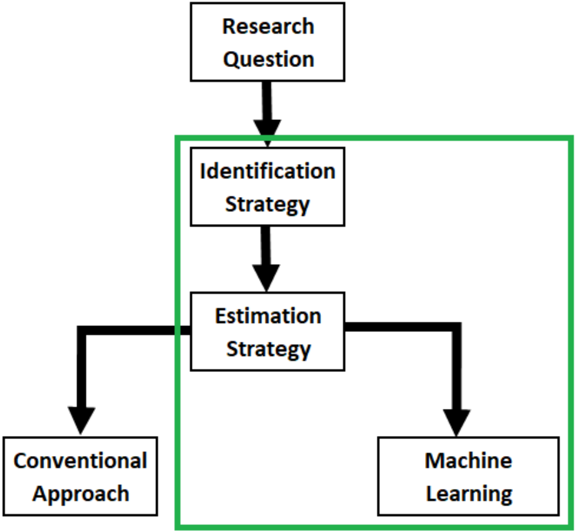
Training of ML Algorithms

Out-of-Sample Mean-Squared-Error (MSE):

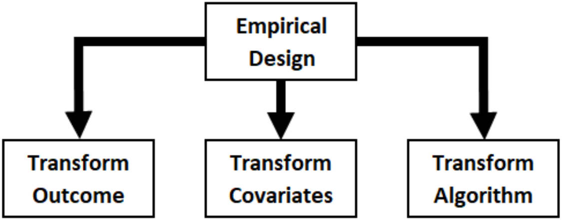
$$MSE_{\hat{\delta}} = E\left[(\hat{\delta} - \delta)^2\right] = \underbrace{E\left[(\hat{\delta} - E[\hat{\delta}])^2\right]}_{\text{Variance}} + \underbrace{E[\hat{\delta} - \delta]^2}_{\text{Squared Bias}}$$

$\rightarrow \delta$ is unobservable

Research Design



Causal Machine Learning (CML) Designs



⇒ Knaus, Lechner, Strittmatter (2018) provide a comparison of all designs.

Potentials of Causal Machine Learning (CML)

Four potential applications of CML:

1. Account for (very) many instruments in IV or Heckit approach (prediction problem, issues with inference).
   References:
   - Belloni, Chen, Chernozhukov, and Hansen (2012)
   - Hansen and Kozbur (2014)
2. Account for confounders, e.g., in matching, IV, or difference-in-difference approaches:
   - ML enables the incorporation of (very) many covariates which can make the exclusion restriction more credible.
   - Some ML approaches make little functional form assumptions.

   Reference:
   - Chernozhukov et al. (2017)

Potentials of Causal Machine Learning (CML)

3. Heterogeneous effects:
   - Principled approach makes it less likely to overlook important heterogeneity.
   - Problems: Issues with interpretability and works only for the low-dimensional case.

   References:
   - Wager and Athey (2018)
   - Chernozhukov, Demirer, Duflo, and Fernéndez-Val (2018)
4. Optimal policy rules (e.g., Bandits):
   - Focus on the (discrete) treatment decision instead on the effect size.

   Reference:
   - Athey and Wager (2019)

Limitations of Causal Machine Learning (CML)

- ML algorithms cannot distinguish between causation and correlation.
  - $\rightarrow$ CML will not select the relevant causal parameters automatically.
  - $\rightarrow$ We have to provide some structure to the CML algorithm.
- CML can estimate causal effects only for a few (usually only one) endogenous variables.
  - $\rightarrow$ We will not obtain the (complete) structural model.
- Identifying assumptions do not change, no matter if we use ML or conventional methods.
- We should resist the temptation to interpret prediction models in a causal way.

Applications of CML Methods

- Knaus (2018) estimates the effects of musical practice on student's skills and selects confounders with ML methods.
- Davis and Heller (2017) investigate the effects of summer jobs on the probability of committing a violent crime.
- Taddy et al. (2016) investigate the heterogeneous effects of A/B-experiments in online-auctions (eBay) on customer responses (experimental study).
- Bertrand et al. (2017) and Knaus, Lechner, and Strittmatter (2020) estimate heterogeneous employment effects of training programmes for unemployed persons.
- Ascarza (2018) targets marketing campaigns.

Supervised vs. Unsupervised Machine Learning

Supervised Machine Learning:

- We observe data on Y and X and want to learn the mapping $\widehat{Y} = \widehat{f}(X)$
- Classification when $\widehat{Y}$ is discrete, regression when $\widehat{Y}$ is continuous

Unsupervised Machine Learning:

- We observe only data on X and want to learn something about its structure
- Clustering: Partition data into homogeneous groups based on X