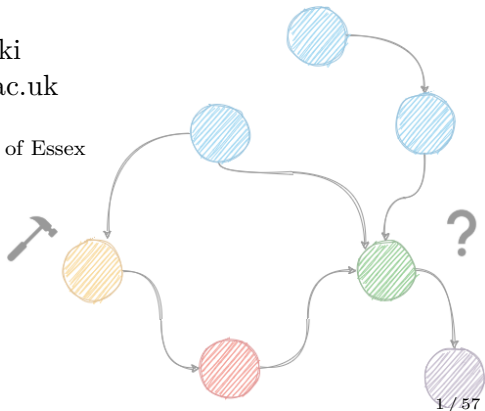


Machine Learning for Causal Inference from Observational Data

Damian Machlanski
d.machlanski@essex.ac.uk

CSEE and MiSoC, University of Essex

July 28th 2022



- ▶ Introduction
- ▶ Motivation
- ▶ Causality
- ▶ Methods
- ▶ Metrics
- ▶ Conclusion

WELCOME!

- ▶ Agenda
 - ▶ Slides: Introduction to Causal Inference
 - ▶ Tutorial: Guided Example with Code
 - ▶ Exercise: Do It Yourself

With some breaks in the middle as necessary.

RESOURCES

► Textbooks

- J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect, 1st ed. USA: Basic Books, Inc., 2018.¹
- J. Pearl, M. Glymour, and N. P. Jewell, Causal Inference in Statistics: A Primer. John Wiley & Sons, 2016.²
- J. Peters, D. Janzing, and B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press, 2017.³

► Online

- Introduction to Causal Inference⁴

¹<http://bayes.cs.ucla.edu/WHY/>

²<http://bayes.cs.ucla.edu/PRIMER/>

³<https://mitpress.mit.edu/books/elements-causal-inference>

⁴<https://www.bradyneal.com/causal-inference-course>

TOOLS

We are going to use the following:

- ▶ Python 3
- ▶ scikit-learn (ML methods)
- ▶ EconML⁵ (CI estimators)
- ▶ The usual ML stack (numpy, pandas, matplotlib)
- ▶ Google Colab

⁵<https://github.com/microsoft/EconML>

A MACHINE LEARNING PERSPECTIVE

We will need the following:

- ▶ Supervised learning - predict y given (X, y) samples
 - ▶ Regression (continuous outcome)
 - ▶ Classification (binary outcome)
- ▶ Basic data exploration
- ▶ Data pre-processing
- ▶ Training and testing
- ▶ Using metrics

You probably know all this by now -> we can do causal inference!

WHY DO I NEED THIS?

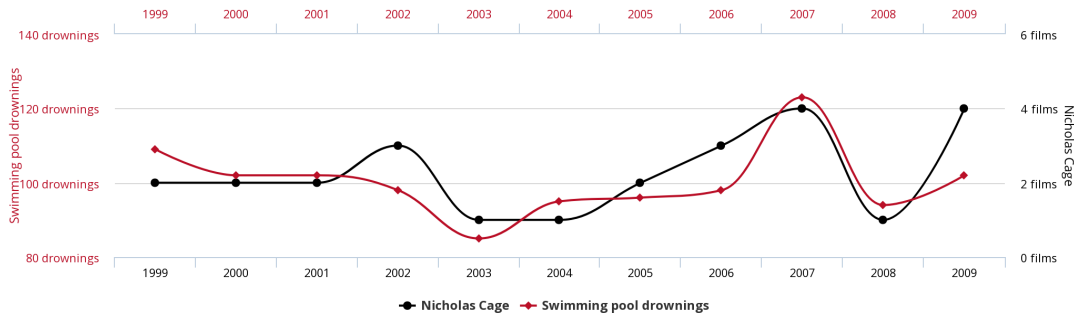
- ▶ Data science is more than just ML
- ▶ It's about **decision making**
- ▶ Associations vs. causal relations
- ▶ *Correlation does not imply causation*
- ▶ Also: biases and shifts within the data that skew the results
- ▶ Wrong conclusions -> bad decisions
- ▶ Complimentary to permutation tests:
 - ▶ PT: Is the effect statistically significant? (yes/no)
 - ▶ CI: How big the effect is? (number)

SPURIOUS CORRELATIONS

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

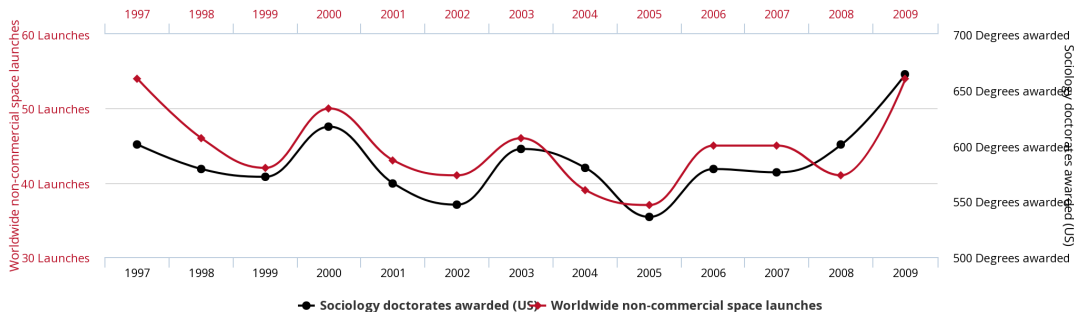


tylervigen.com

Credit: <https://www.tylervigen.com/spurious-correlations>

SPURIOUS CORRELATIONS (2)

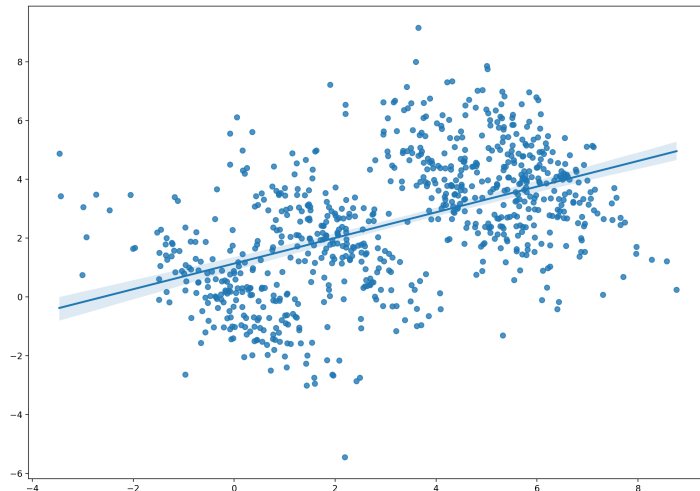
Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)



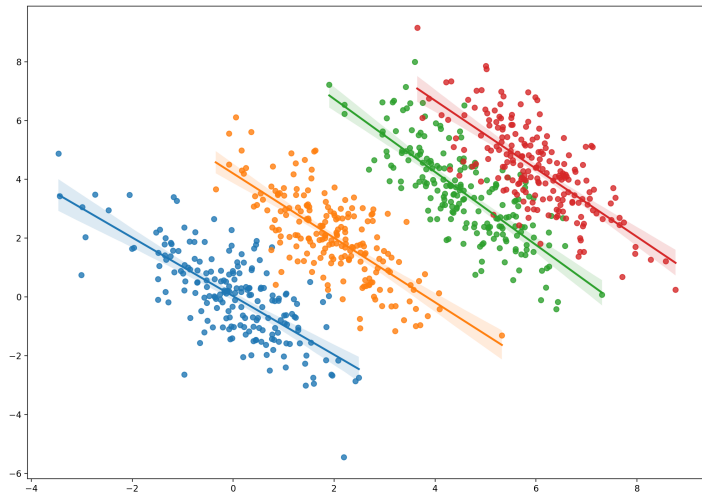
tylervigen.com

Credit: <https://www.tylervigen.com/spurious-correlations>

SIMPSON'S PARADOX

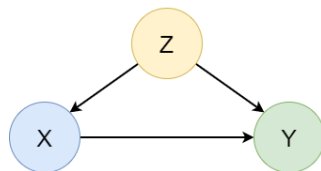
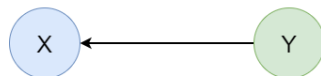
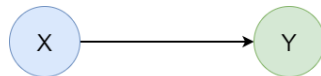


SIMPSON'S PARADOX



TAKEAWAY

- ▶ We need to think about the causal links within the data (causal graphs).
- ▶ Cause and effect
- ▶ Question: As we *change* the cause, how does the effect *change*?



PROBLEM SETTING

We want to estimate the *causal effect* of treatment T on outcome Y

- ▶ What benefits accrue if we intervene to change T ?
- ▶ Treatment must be modifiable
- ▶ We observe only one outcome per each individual

Ideal scenario:

1. Assume state S_0
2. Apply the treatment ($t = 1$)
3. Observe the outcome (Y_1)
4. Reset the state to S_0 (steps 2. and 3. didn't happen)
5. Do not apply the treatment ($t = 0$)
6. Observe the outcome (Y_0)
7. Compare the outcomes Y_1 and Y_0 to get the causal effect

REAL-LIFE EXAMPLE

- ▶ My headache went away after I had taken the aspirin (Y_1)
- ▶ Would the headache have gone away without taking the aspirin? ($Y_0 = ?$)
- ▶ We cannot go back in time and test the alternative!
- ▶ Cannot reset the state -> cannot compare the outcomes -> no effect
- ▶ Test more people and measure the average outcome?

MORE EXAMPLES

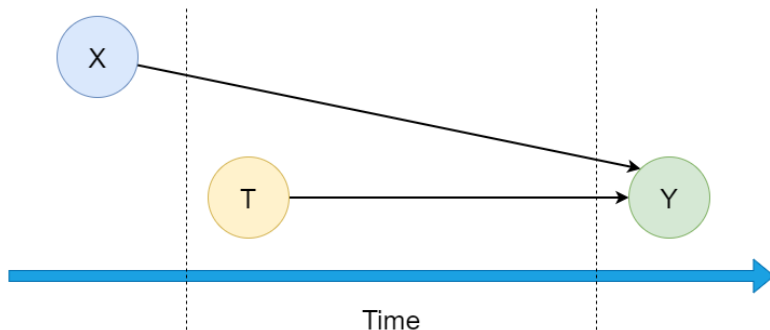
- ▶ Developing a new vaccine
- ▶ Government policy
- ▶ Recommending the best treatment for a specific patient

It's about finding out how a specific action affects a system of interest.

- ▶ Action == intervention (something we change)
- ▶ System == the very thing we study (group of people, physical objects, etc.)
- ▶ Outcome == system's characteristic of interest (response)
- ▶ Effect == difference between outcomes

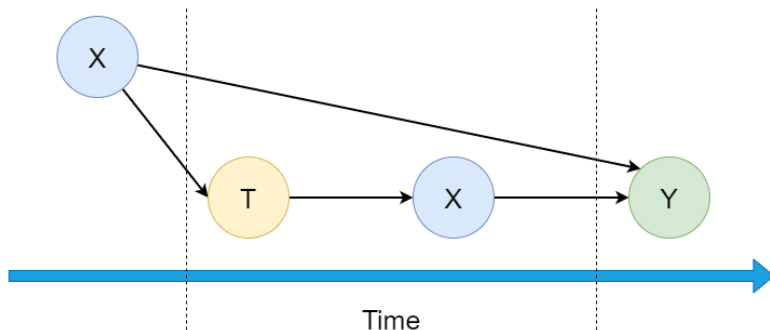
RANDOMISED CONTROLLED TRIALS

- ▶ Data from controlled experiments
- ▶ Randomised T - people assigned $T = 0$ (control) or $T = 1$ (treated)
- ▶ This mimicks observing alternative reality
- ▶ Record background characteristics as $X = [X_1, X_2, \dots, X_n]$
- ▶ Can be expensive or even unfeasible (e.g. smoking)



OBSERVATIONAL DATA

- ▶ Passively collected data (non-experimental)
- ▶ Abundant nowadays
- ▶ Quasi-experimental study
- ▶ Keep only X recorded before Y (discard other)
- ▶ Lack of randomisation and control (imbalances)



ML PERSPECTIVE

- ▶ Correlation vs. causation
- ▶ Outliers - different meaning
- ▶ Imbalanced data (not just Y)
- ▶ Out-of-distribution (OOD) generalisation
- ▶ ML vs. CI:
 - ▶ ML: predict Y given (X, Y) samples
 - ▶ CI: predict **effects** given (X, Y) samples

MORE ON ML vs. CI

ML

- ▶ Train on (X, Y) samples
- ▶ Predict Y given X test samples
- ▶ Assumes the same distribution of training and testing samples

CI

- ▶ Train on (X, T, Y) samples
- ▶ Predict Y for (X, T) and $(X, \mathbf{1-T})$
- ▶ $(X, 1-T)$: predict the outcomes we haven't observed
- ▶ Treated ($t = 1$) and control ($t = 0$) groups often have different distributions
- ▶ We learn from one distribution, but make predictions for a different one!
- ▶ The usual IID assumption no longer applies here

FUNDAMENTALS

$$Effect = Y_1 - Y_0$$

#	X_1	X_2	X_3	T	Y_0	Y_1
1	1.397	0.996	0	1	?	4.771
2	0.269	0.196	1	0	2.956	?
3	1.051	1.795	1	1	?	4.164
4	0.662	0.196	0	1	?	6.172
5	0.856	1.795	1	0	7.834	?

Observed and unobserved outcomes are **factuals** and **counterfactuals** respectively.

Missing counterfactuals: This is known as the fundamental problem of causal inference.

We cannot *observe* the difference, but we can **approximate** it.

TREATMENT EFFECT

Let us define the **true** outcome $\mathcal{Y}_t^{(i)}$ of individual (i) that received treatment $t \in \{0, 1\}$. The Individual Treatment Effect (ITE) is then defined as follows:

$$ITE^{(i)} = \mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}$$

The Average Treatment Effect (ATE) builds on ITE:

$$ATE = \mathbb{E}[ITE]$$

Note: empirical (sample) ATE is the mean of ITEs.

TREATMENT EFFECT - ITE EXAMPLE

We are given the outcomes Y for both the treated ($t = 1$) and control ($t = 0$) case, where $Y_1 = 3$ and $Y_0 = 2$.

What is the value of ITE?

TREATMENT EFFECT - ITE EXAMPLE (2)

We are given the outcomes Y for both the treated ($t = 1$) and control ($t = 0$) case, where $Y_1 = 3$ and $Y_0 = 2$.

What is the value of ITE?

$$ITE^{(i)} = y_1^{(i)} - y_0^{(i)}$$

$$ITE = 3 - 2 = 1$$

TREATMENT EFFECT - ATE EXAMPLE

We are given the following data:

- ▶ $Y_0 \in \{2, 3, 1\}$
- ▶ $Y_1 \in \{3, 4, 2\}$

What is the value of ATE?

TREATMENT EFFECT - ATE EXAMPLE (2)

We are given the following data:

- ▶ $Y_0 \in \{2, 3, 1\}$
- ▶ $Y_1 \in \{3, 4, 2\}$

What is the value of ATE?

$$ATE = \mathbb{E}[ITE]$$

$$ITE^{(0)} = 3 - 2 = 1$$

$$ITE^{(1)} = 4 - 3 = 1$$

$$ITE^{(2)} = 2 - 1 = 1$$

$$ATE = \frac{ITE^{(0)} + ITE^{(1)} + ITE^{(2)}}{3} = \frac{1 + 1 + 1}{3} = \frac{3}{3} = 1$$

TREATMENT EFFECT - CATE

A more general way of defining effects is through conditioning:

$$CATE = \mathbb{E}[\mathcal{Y}_1|X = x] - \mathbb{E}[\mathcal{Y}_0|X = x]$$

Which stands for Conditional Average Treatment Effect.

Note the two previous effects are special cases of CATE (ATE: $x = \emptyset$, ITE: unique x).

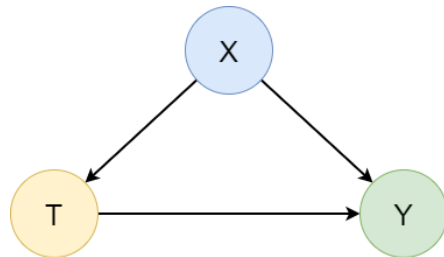
You will likely see CATE estimators in the literature and CI packages.

ASSUMPTIONS

- ▶ Ignorability:
 - ▶ No hidden confounders (we observe everything)
- ▶ All background covariates X happened *before* the outcome Y
- ▶ Modifiable treatment T
- ▶ Stable Unit Treatment Value Assumption (SUTVA):
 - ▶ No interference between units
 - ▶ Consistent treatment (different versions disallowed)

ASSUMPTIONS (2)

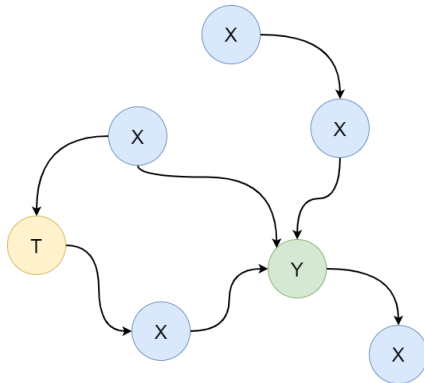
- ▶ Most CI estimators assume the *triangle* graph



- ▶ This is a very simplistic view of the world
- ▶ Actual reality can be much more complex

ASSUMPTIONS (3)

- ▶ Can we infer graphs from data?
- ▶ Causal discovery



ONTO THE METHODS

We know the theory. Now, let's do some modelling!

MODERN APPROACHES

Mostly regression and classification (classic ML), but combined in a smart way.

- ▶ Recent surveys on modern causal inference methods ^{6 7}
- ▶ Most popular:
 - ▶ Inverse Propensity Weighting (IPW)
 - ▶ Doubly-Robust
 - ▶ Double/Debiased Machine Learning
 - ▶ Causal Forests
 - ▶ Meta-Learners
 - ▶ Multiple based on neural networks (very advanced)

Too many to discuss here, but we will learn some common principles.

We will start with a simple regression, add IPW, and conclude with Meta-Learners.

⁶<https://dl.acm.org/doi/10.1145/3397269>

⁷<https://arxiv.org/abs/2002.02770>

S-LEARNER

We want to estimate

$$\mu(t, x) = \mathbb{E}[\mathcal{Y} | X = x, T = t]$$

1. Obtain $\hat{\mu}(t, x)$ estimator.
2. Predict ITE as

$$\widehat{ITE}(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x)$$

- ▶ *Single* model approach
- ▶ Allows heterogenous treatment effects
- ▶ Can be biased (next slide)

S-LEARNER - CODE

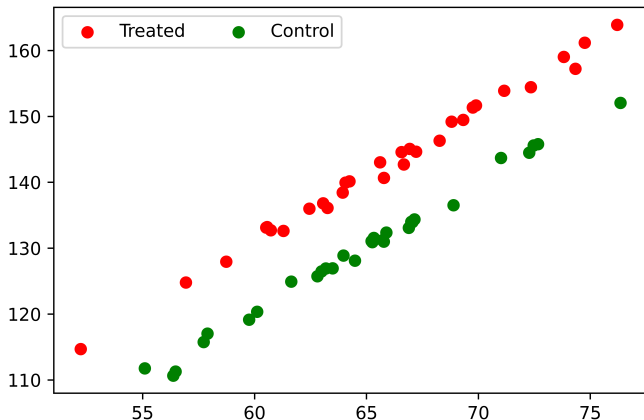
```
lr = LinearRegression()

# input: [X, T], target: Y
lr.fit(np.concatenate([x_train, t_train], axis=1), y_train)

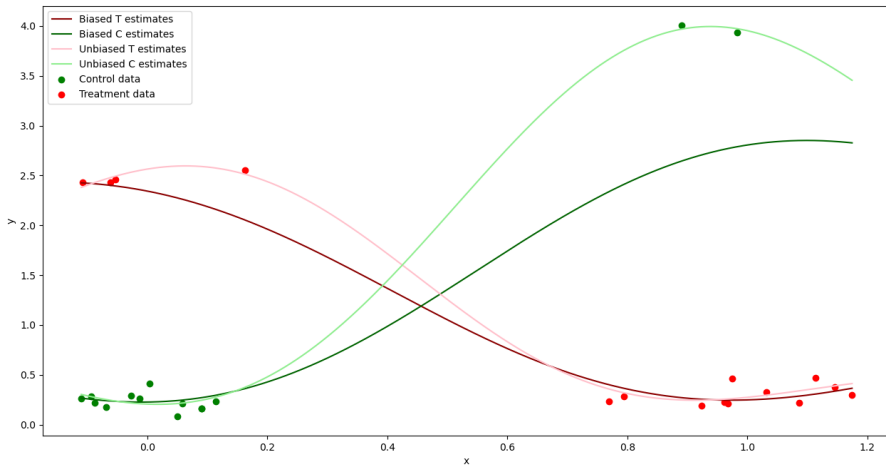
# predict Y0 given [X, 0] - set T=0
y0_pred = lr.predict(np.concatenate([x_test, np.zeros_like(t_test)], axis=1))
# predict Y1 given [X, 1] - set T=1
y1_pred = lr.predict(np.concatenate([x_test, np.ones_like(t_test)], axis=1))

# effect = y1 - y0
effect_pred = y1_pred - y0_pred
```

WHEN IT WORKS



BIASED ESTIMATORS



PROPENSITY SCORE

$$e(x) = P(t_i = 1 | x_i = x)$$

- ▶ Probability of a unit i receiving the treatment ($T = 1$)
- ▶ For discrete treatments, this is a classification problem
- ▶ Binary classification in most cases as $t \in \{0, 1\}$
- ▶ We denote $\hat{e}(x)$ as our estimation

IPW ESTIMATOR

Using the propensity score $\hat{e}(x)$, we can obtain the following weights

$$w_i = \frac{t_i}{\hat{e}(x_i)} + \frac{1 - t_i}{1 - \hat{e}(x_i)}$$

- ▶ These are called Inverse Propensity Weights (IPW)
- ▶ Use the weights to perform **weighted** regression
- ▶ Similar to S-Learner, but combines regression and classification
- ▶ Sample importance (pay attention to scarce data points)
- ▶ Either $\hat{e}(x)$ or $\hat{\mu}(x)$ can still have bias (misspecification)
- ▶ Doubly-Robust method attempts to address that

IPW ESTIMATOR - CODE

```
clf = LogisticRegression()
weights = get_ps_weights(clf, x_train, t_train)

lr = LinearRegression()

# input: [X, T], target: Y
lr.fit(np.concatenate([x_train, t_train], axis=1), y_train, sample_weight=weights)

# ...
```

T-LEARNER

- ▶ Treated and control distributions are often different
- ▶ Solution: fit *two* separate regressors

$$\mu_1(x) = \mathbb{E}[\mathcal{Y}|X = x, T = 1]$$

$$\mu_0(x) = \mathbb{E}[\mathcal{Y}|X = x, T = 0]$$

1. Learn $\mu_1(x)$ from treated units, obtain $\hat{\mu}_1(x)$.
2. Learn $\mu_0(x)$ from control units, obtain $\hat{\mu}_0(x)$.
3. Predict ITE as

$$\widehat{ITE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

T-LEARNER - CODE

```
m0 = LinearRegression()
m1 = LinearRegression()

t0_idx = (t_train == 0).flatten()
t1_idx = (t_train == 1).flatten()

# train on control units
m0.fit(x_train[t0_idx], y_train[t0_idx])
# train on treated units
m1.fit(x_train[t1_idx], y_train[t1_idx])

y0_pred = m0.predict(x_test)
y1_pred = m1.predict(x_test)

effect_pred = y1_pred - y0_pred
```


T-LEARNER - CODE (2)

```
t1 = TLearner(models=LinearRegression())  
  
t1.fit(y_train, t_train, X=x_train)  
  
effect_pred = t1.effect(x_test)
```

X-LEARNER

A hybrid of the previous approaches (details here⁸). There are three main stages.

1. Learn treated and control separately (same as T-Learner).
2. Predict and learn *imputed* effects (mix of Y_f and Y_{cf}).
3. Learn a propensity score function.

The final treatment effect estimate is a weighted average of the two estimates from Stage 2:

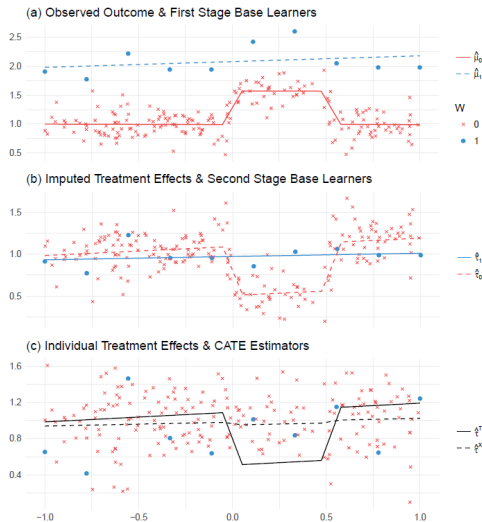
$$\hat{\tau}(x) = \hat{e}(x)\hat{\tau}_0(x) + (1 - \hat{e}(x))\hat{\tau}_1(x)$$

⁸<http://arxiv.org/abs/1706.03461>

X-LEARNER - CODE

```
xl = XLearner(models=LinearRegression(), propensity_model=LogisticRegression())  
  
xl.fit(y_train, t_train, X=x_train)  
  
effect_pred = xl.effect(x_test)
```

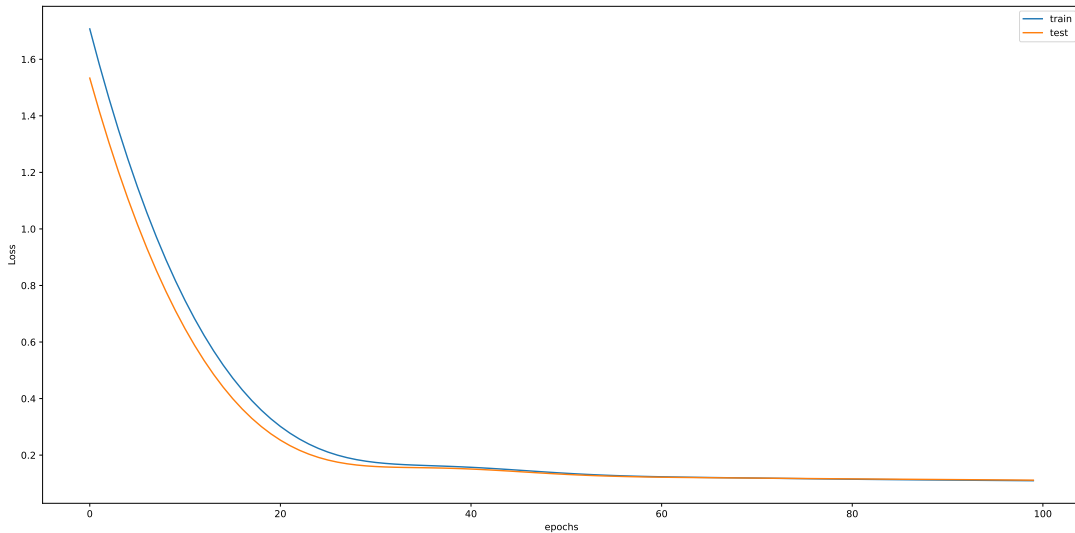
X-LEARNER - INTUITION



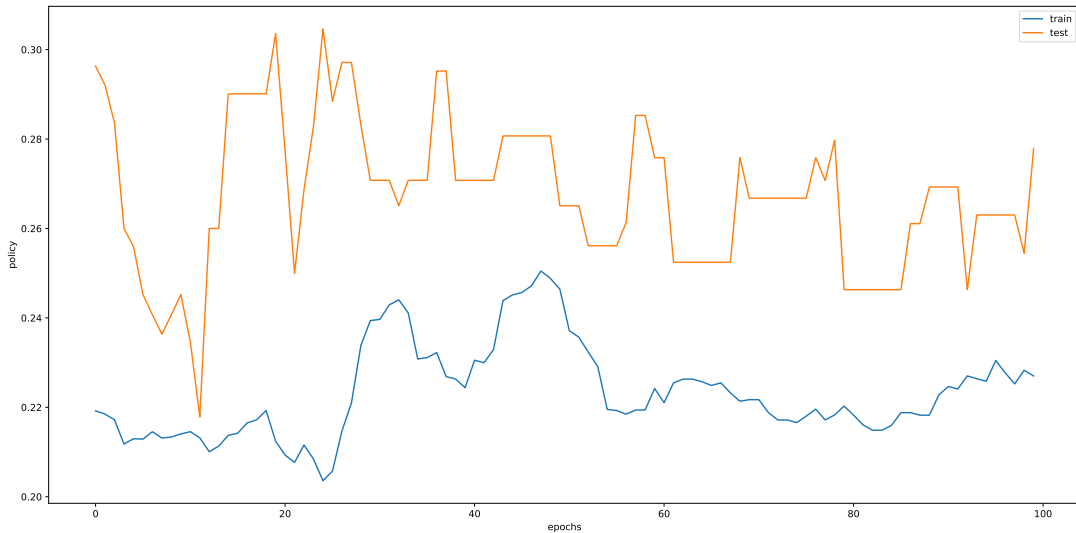
EVALUATION

- ▶ We have predicted some effects.
- ▶ But are they accurate?
- ▶ How good our model is at predicting effects?
- ▶ Can we use the usual metrics like MSE?

MSE



POLICY RISK



ERROR ON OUTCOMES VS. EFFECTS

- ▶ Predicting accurate *outcomes* Y (MSE) is only half of the problem
- ▶ However, the priority is to predict accurate **effects**
- ▶ Thus, we need to measure the amount of error (ϵ) or risk (\mathcal{R}) introduced by a model with respect to predicted effects

Examples:

- ▶ ϵ_{ATE}
- ▶ ϵ_{PEHE}
- ▶ ϵ_{ATT}
- ▶ \mathcal{R}_{pol}

PREDICTIONS

Let us denote $\hat{y}_t^{(i)}$ as **predicted** outcome for individual (i) that received treatment t . Then, our predicted ITE and ATE can be written as:

$$\widehat{ITE}^{(i)} = \hat{y}_1^{(i)} - \hat{y}_0^{(i)}$$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \widehat{ITE}^{(i)}$$

MEASURING ERRORS

This allows us to define the following measurement errors:

$$\epsilon_{PEHE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\widehat{ITE}^{(i)} - ITE^{(i)})^2}$$

$$\epsilon_{ATE} = |\widehat{ATE} - ATE|$$

Where *PEHE* stands for Precision in Estimation of Heterogeneous Effect, and which essentially is a Root Mean Squared Error (RMSE) between predicted and true ITEs.

BENCHMARK DATASETS

Semi-simulated data or combinations of experimental and observational datasets. We use metrics depending on what kind of information we have access to (true effects/counterfactuals).

The main purpose is to test performance of causal estimators. The focus is on methods, not data.

Some well-established causal inference datasets:

- ▶ IHDP
- ▶ Jobs
- ▶ News
- ▶ Twins
- ▶ ACIC challenges

TYPES OF METRICS

With effect/counterfactuals

- ▶ ϵ_{ATE}
- ▶ ϵ_{PEHE}
- ▶ Datasets with simulated outcomes
- ▶ (it's unnatural to observe both outcomes!)

Without effect/counterfactuals

- ▶ ϵ_{ATT} (ATE on the Treated)
- ▶ \mathcal{R}_{pol} (Policy Risk)
- ▶ Datasets closer to reality
- ▶ Either purely observational or mixed with RCTs

THERE IS MORE

- ▶ We just scratched the surface here
- ▶ Causal discovery (inferring graphs from data) - big topic on its own
- ▶ Estimating causal effects vs. recommending treatments⁹
- ▶ Other methods
 - ▶ Instrumental variables
 - ▶ Relaxing the common assumptions
 - ▶ Trees, neural networks, policy learners
- ▶ Front-door and back-door adjustments
- ▶ Handling colliders, confounders, feature selection
- ▶ ...

⁹<http://arxiv.org/abs/2104.04103>

SUMMARY

- ▶ Causal inference is about estimating causal effects
 - ▶ For instance, measure the effectiveness of a treatment
- ▶ RCTs are the most reliable source of data, but can be unfeasible to obtain
- ▶ Non-experimental data are a great alternative, but can be *biased*
- ▶ Most methods are about finding *unbiased* estimators
- ▶ Machine Learning and Causal Inference can be both mutually beneficial
 - ▶ ML delivers better CI estimators
 - ▶ CI helps ML with OOD generalisation
- ▶ Assumptions and graphs are important and must be considered in applications

ACKNOWLEDGEMENTS

This lecture builds heavily on the materials from *Introduction to Machine Learning for Causal Analysis Using Observational Data* online course, delivered on June 22-23 2021 by Damian Machlanski, Dr Spyros Samothrakis and Professor Paul Clarke.

REFERENCES

- ▶ J. M. Robins, A. Rotnitzky, and L. P. Zhao, ‘Estimation of Regression Coefficients When Some Regressors are not Always Observed’, Journal of the American Statistical Association, vol. 89, no. 427, pp. 846–866, Sep. 1994.
- ▶ U. Shalit, F. D. Johansson, and D. Sontag, ‘Estimating individual treatment effect: generalization bounds and algorithms’, in International Conference on Machine Learning, Jul. 2017, pp. 3076–3085.
- ▶ V. Chernozhukov et al., ‘Double/debiased machine learning for treatment and structural parameters’, The Econometrics Journal, vol. 21, no. 1, pp. C1–C68, Feb. 2018.
- ▶ S. Wager and S. Athey, ‘Estimation and Inference of Heterogeneous Treatment Effects using Random Forests’, Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228–1242, Jul. 2018.
- ▶ S. R. Künnel, J. S. Sekhon, P. J. Bickel, and B. Yu, ‘Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning’, Proc Natl Acad Sci USA, vol. 116, no. 10, pp. 4156–4165, Mar. 2019.
- ▶ R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, ‘A Survey of Learning Causality with Data: Problems and Methods’, ACM Comput. Surv., vol. 53, no. 4, p. 75:1–75:37, Jul. 2020.
- ▶ L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, ‘A Survey on Causal Inference’, arXiv:2002.02770 [cs, stat], Feb. 2020.

WHAT'S NEXT?

- ▶ Onto the practical parts
 - ▶ Tutorial
 - ▶ Predict ATE and measure ϵ_{ATE}
 - ▶ S-Learner, IPW and X-Learner
 - ▶ Random Forest as base regressors and classifiers
 - ▶ Exercise - IHDP
 - ▶ Predict ITE and ATE
 - ▶ Measure ϵ_{PEHE} and ϵ_{ATE}
 - ▶ Exercise - JOBS (optional)
 - ▶ Predict ATT and Policy
 - ▶ Measure ϵ_{ATT} and \mathcal{R}_{pol}
- ▶ Short break?