Machine Learning for econometrics

Causal perspective

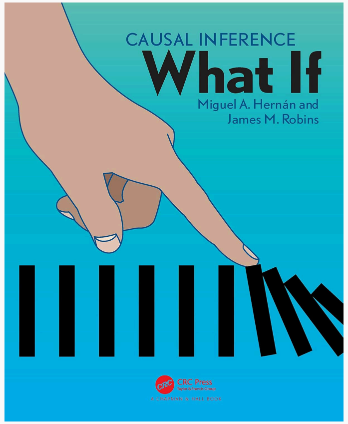
Matthieu Doutreligne

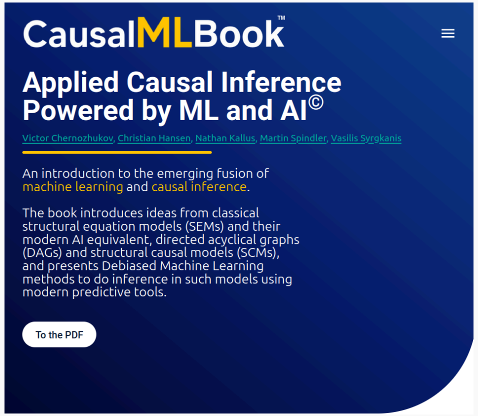
January 10, 2025

Table of contents

Introduction

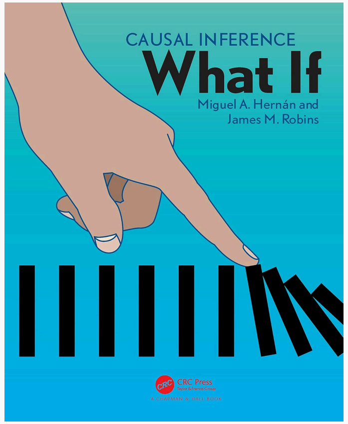
Causal inference: subfield of statistics dealing with "why questions"

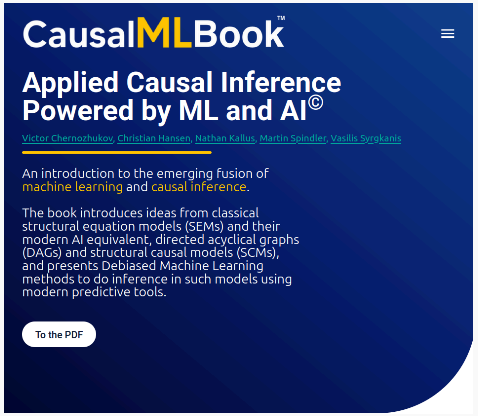




At the center of epidemiology (Hernan & Robins, 2020), econometrics (Chernozhukov et al., 2024), social sciences,

Causal inference: subfield of statistics dealing with "why questions"

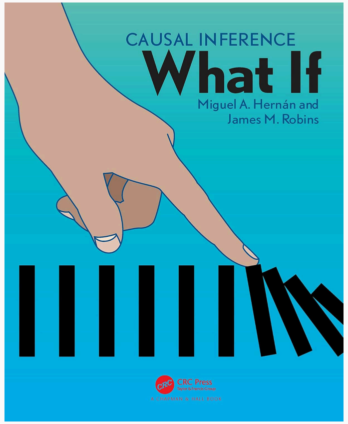


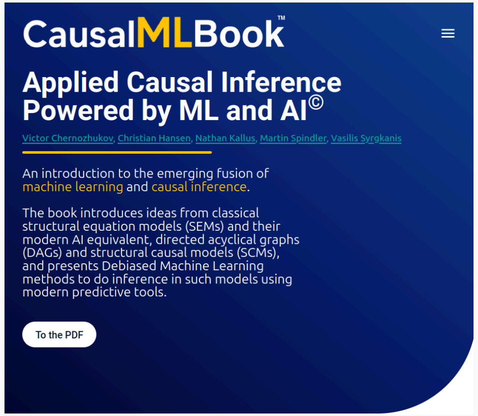


At the center of epidemiology (Hernan & Robins, 2020), econometrics (Chernozhukov, Hansen, Kallus, Spindler, & Syrgkanis, 2024), social sciences, machine learning…

Now, bridging with machine learning (Kaddour et al., 2022) : Fairness, reinforcement learning, causal discovery, causal inference for LLM, causal representations…

Causal inference: subfield of statistics dealing with "why questions"





At the center of epidemiology (Hernan & Robins, 2020), econometrics (Chernozhukov, Hansen, Kallus, Spindler, & Syrgkanis, 2024), social sciences,

This course:

- Basis of causal inference using ML approaches (semi-parametric),
- Inspiration from epidemiology,
- Application in econometrics.

What is a "why question"?

- Economics: How does supply and demand (causally) depend on price?
- Policy: Are job training programmes actually effective?
- Epidemiology: How does this threatment affect the patient's health?
- Public health : Is this prevention campaign effective?
- Psychology: What is the effect of family structure on children's outcome?
- Sociology: What is the effect of social media on political opinions?

This is different from a predictive question

- What will be the weather tomorrow?
- What will be the outcome of the next election?
- How many people will get infected by flue next season?
- What is the cardio-vacular risk of this patient?
- How much will the price of a stock be tomorrow?

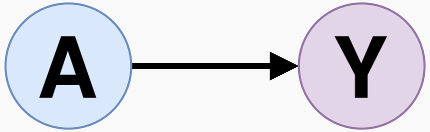
Why is prediction different from causation? (1/2)

Prediction (most part of ML): What usually happens in a given situation?

Why is prediction different from causation? (1/2)

Prediction (most part of ML): What usually happens in a given situation?

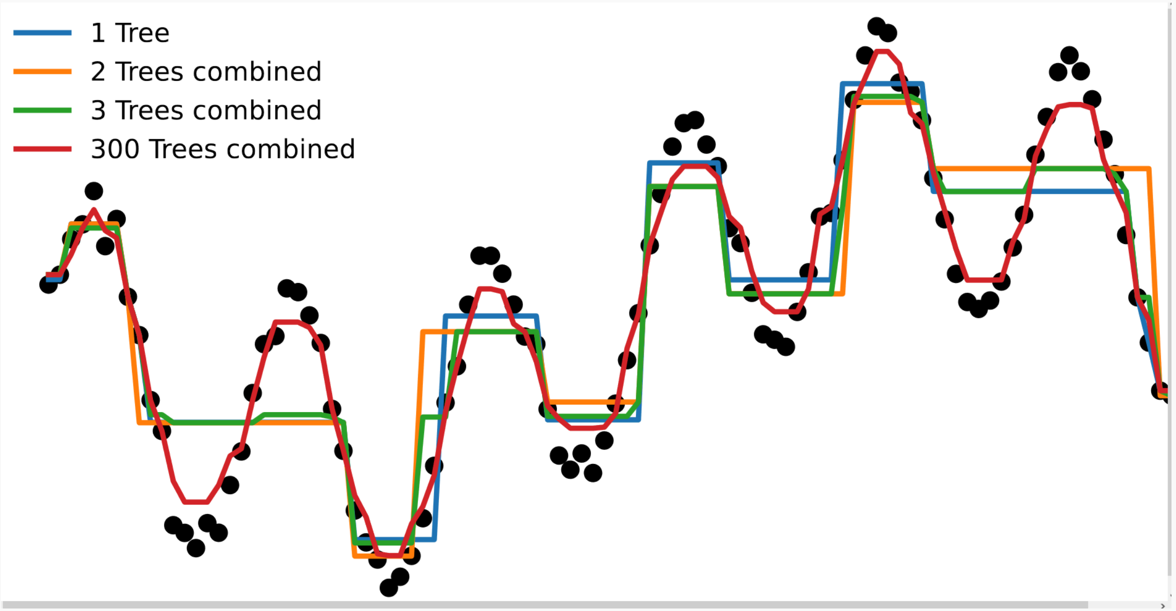
Assumption Train and test data are drawn from the same distribution.



Prediction models (X, Y)

Machine learning is (basically) pattern matching

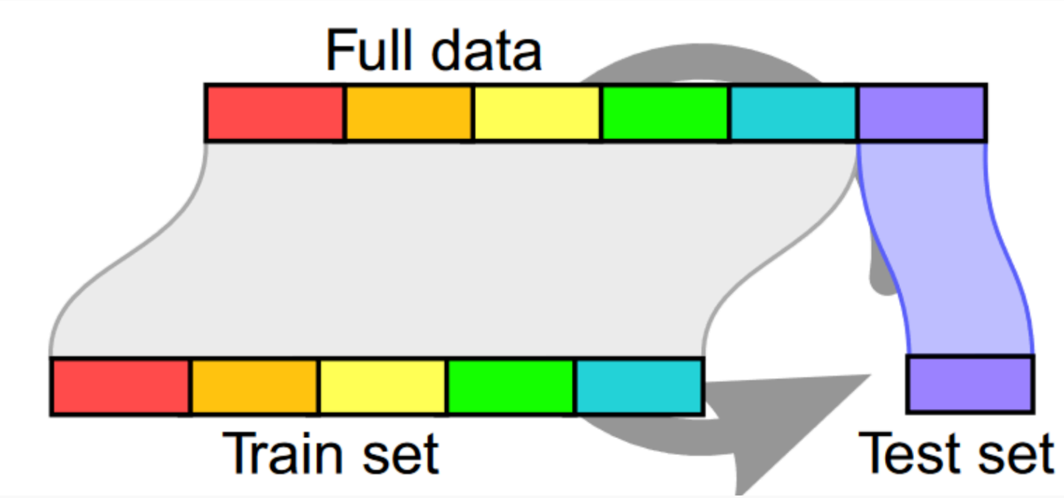
Find an estimator $f : x \to y$ that approximates the true value of y so that $f(x) \approx y$



Boosted trees : iterative ensemble of decision trees

Machine learning is pattern matching that generalizes to new data

Select models based on their ability to generalize to new data : (train, test) splits and cross validation (Stone, 1974).



“Cross validation” (Varoquaux et al., 2017)

Leverages complex data structures

- Images: Image classification with deep convolutional neural networks (Krizhevsky et al., 2012)

Leverages complex data structures

- Speech-to-text: Towards end-to-end speech recognition with recurrent neural networks (Graves & Jaitly, 2014)

Machine learning is great for prediction

Leverages complex data structures

- Text: Attention is all you need (Vaswani, 2017)

Machine learning is not driven by causal mechanisms

- For example people that go to the hospital die more than people who do not[1]:
  - Naive data analysis might conclude that hospitals are bad for health.

  The fallacy is that we are comparing different populations: people who go to the hospital typically have a worse baseline health than people who do not.

This is a confounding factor: a variable that influences both the treatment and the outcome.

[1]Example from https://inria.github.io/scikit-learn-mooc/concluding_remarks.html?highlight=causality

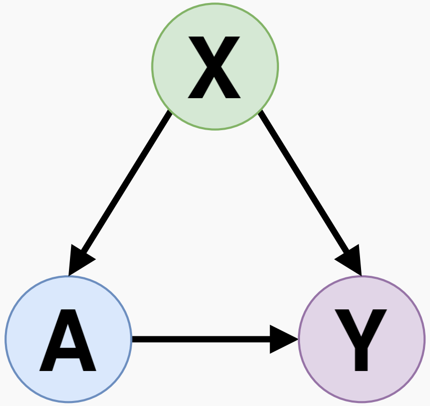
Why is prediction different from causation? (2/2)

Causal inference (most part of economists) : What would happen if we changed the system ie. under an intervention?

Why is prediction different from causation? (2/2)

Causal inference (most part of economists) : What would happen if we changed the system ie. under an intervention?

Assumption: No unmeasured variables influencing both treatment and outcome $\rightarrow$ confounders.



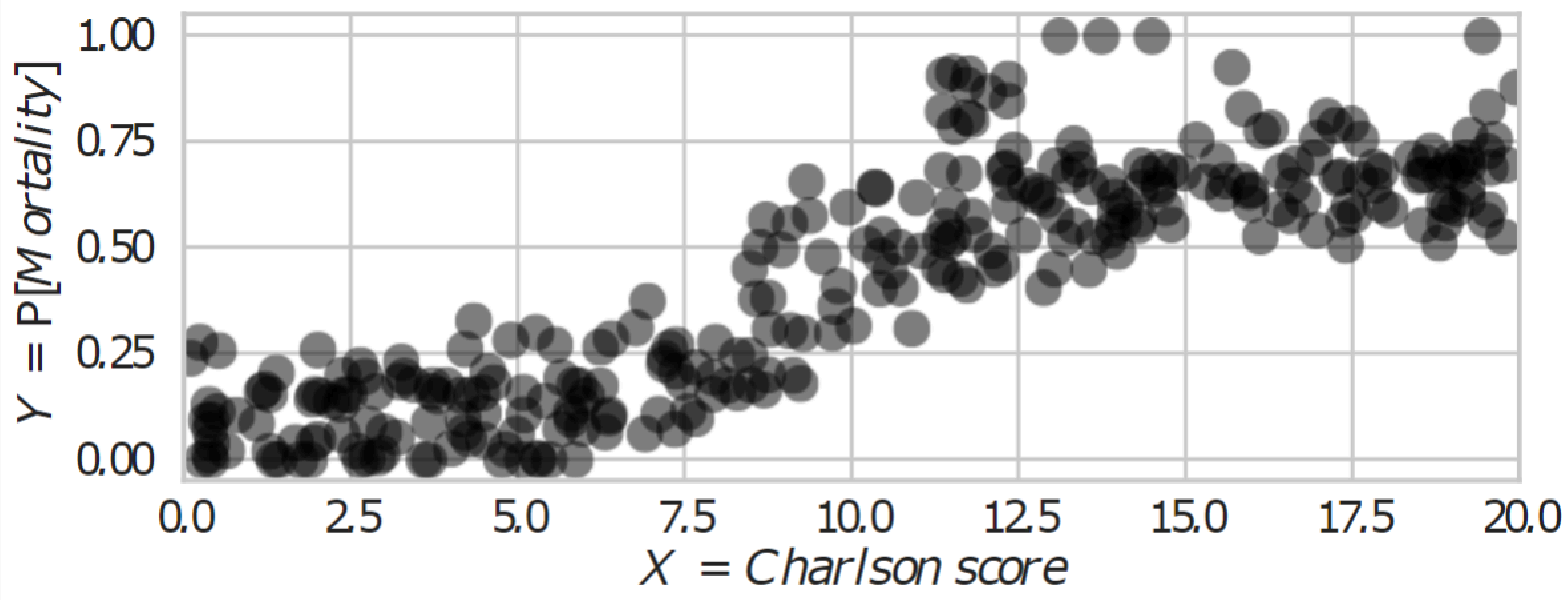
Causal inference models
$(X, A, Y(A=1), Y(A=0))$
the covariate shift between treated and control units.

Illustration of the fundamental problem of causal inference

Consider an example from epidemiology:

- Population: patients experiencing a stroke
- Intervention $A = 1$: patients had access to a MRI scan **in less than 3 hours** after the first symptoms
- Comparator $A = 0$: patients had access to a MRI scan **in more than 3 hours** after the first symptoms
- $Y = \mathbb{P}[\text{Mortality}]$: the mortality at 7 days
- $X = \mathbb{P}[\text{Charlson score}]$: a comorbidity index summarizing the overall health state of the patient. Higher is bad for the patient.

Draw a population sample without treatment status



Draw a population sample with treatment status

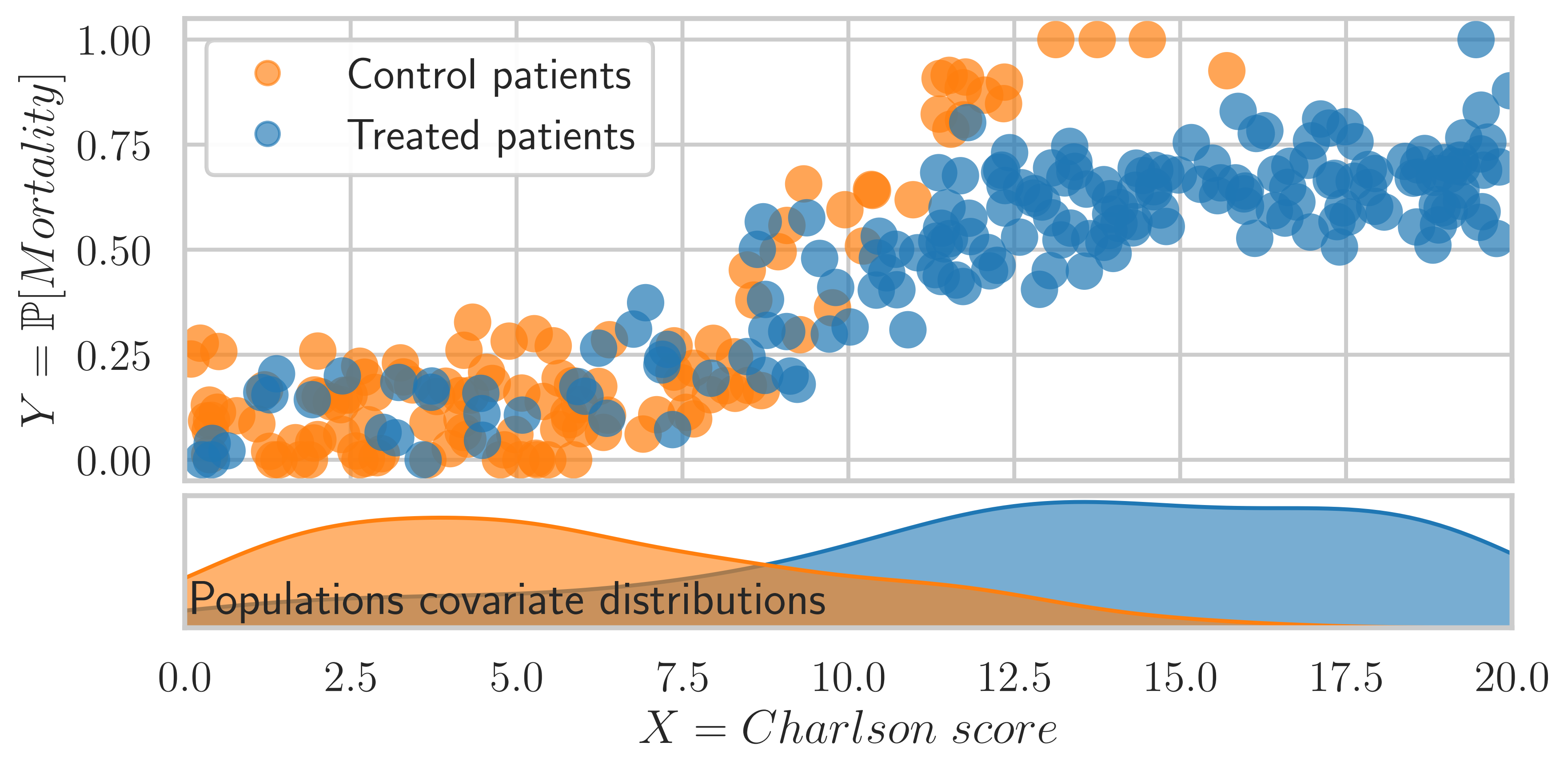
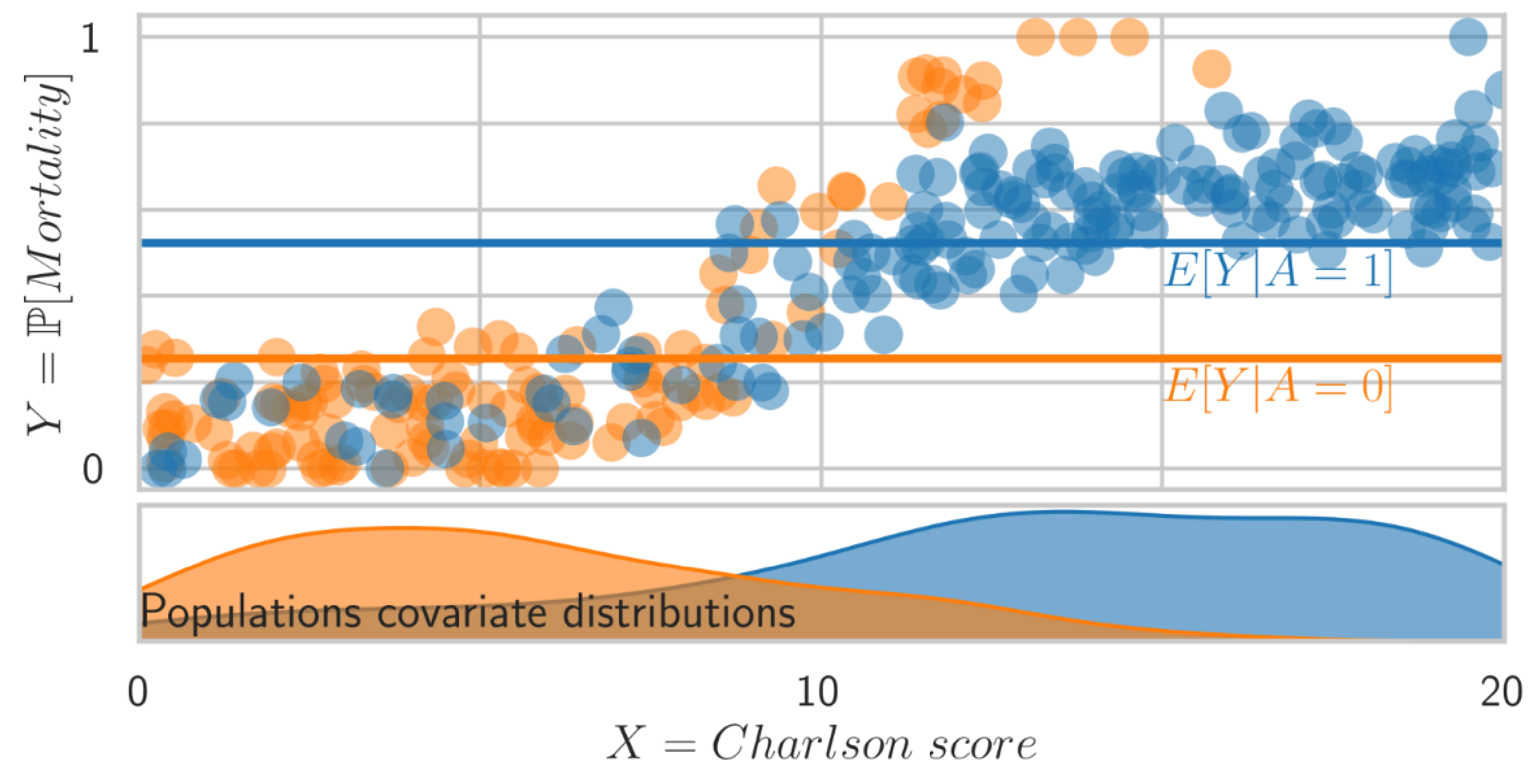


Illustration: observational data, a naive solution

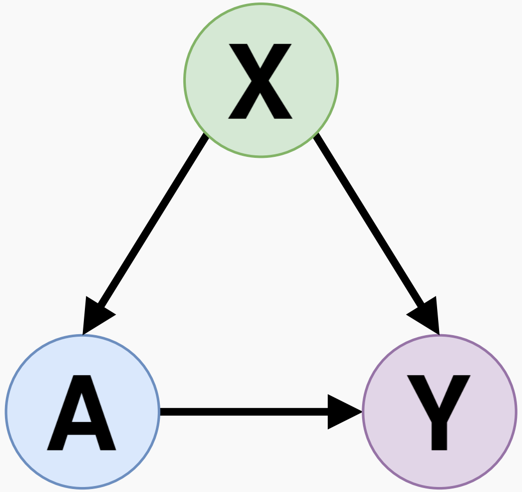
Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$

Illustration: observational data, a naive solution

Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$



Observational data

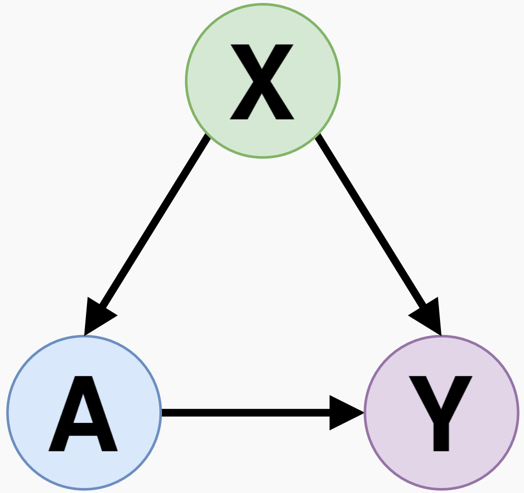


$$Y(1), Y(0) \not\perp\!\!\!\perp A$$

Intervention is not random

(with respect to the confounders)

Observational data

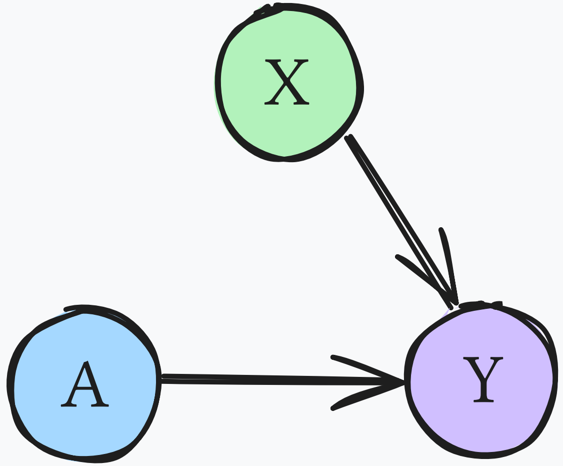


$$Y(1), Y(0) \not\perp\!\!\!\perp A$$

Intervention is not random

(with respect to the confounders)

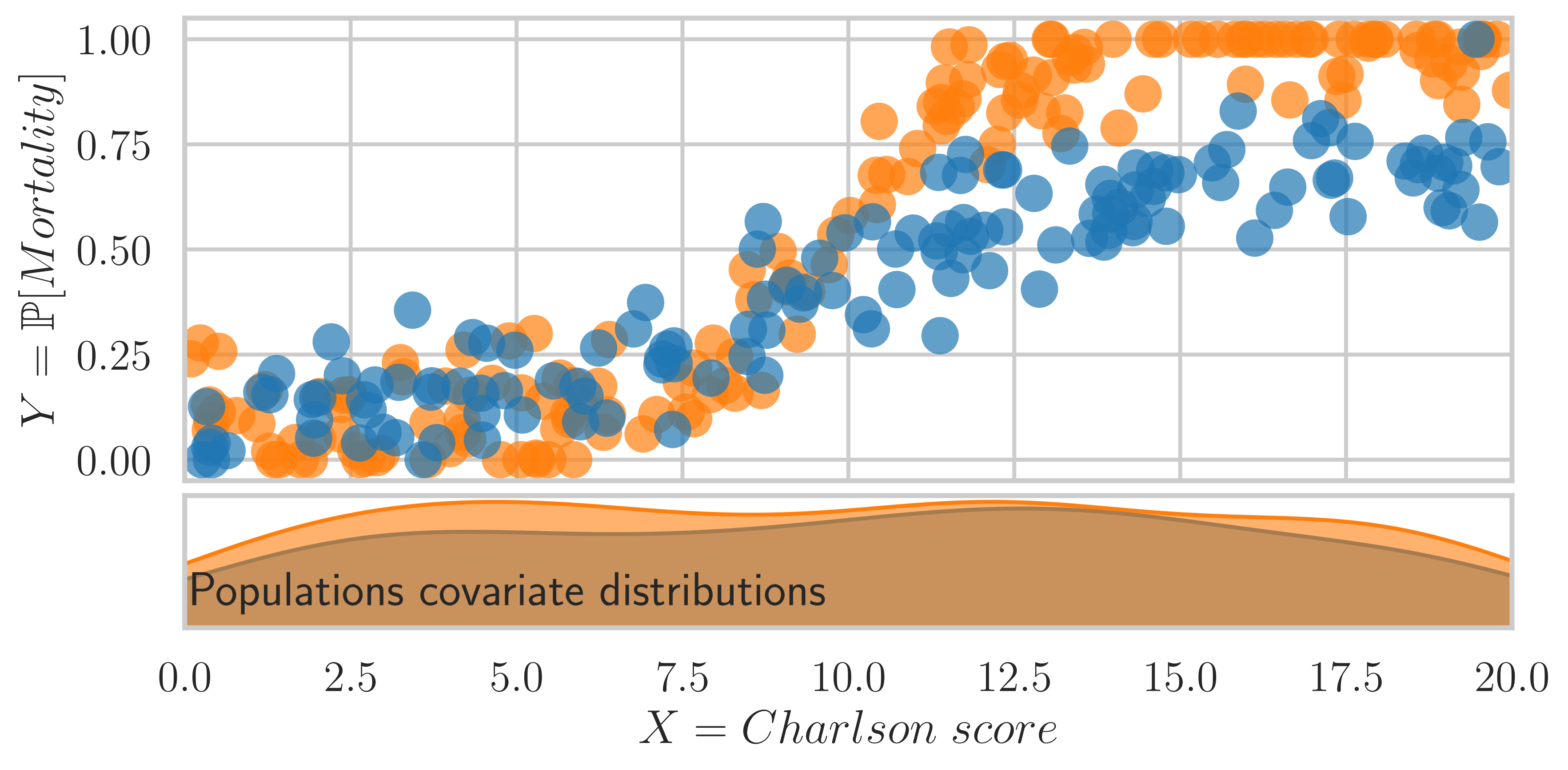
RCT data



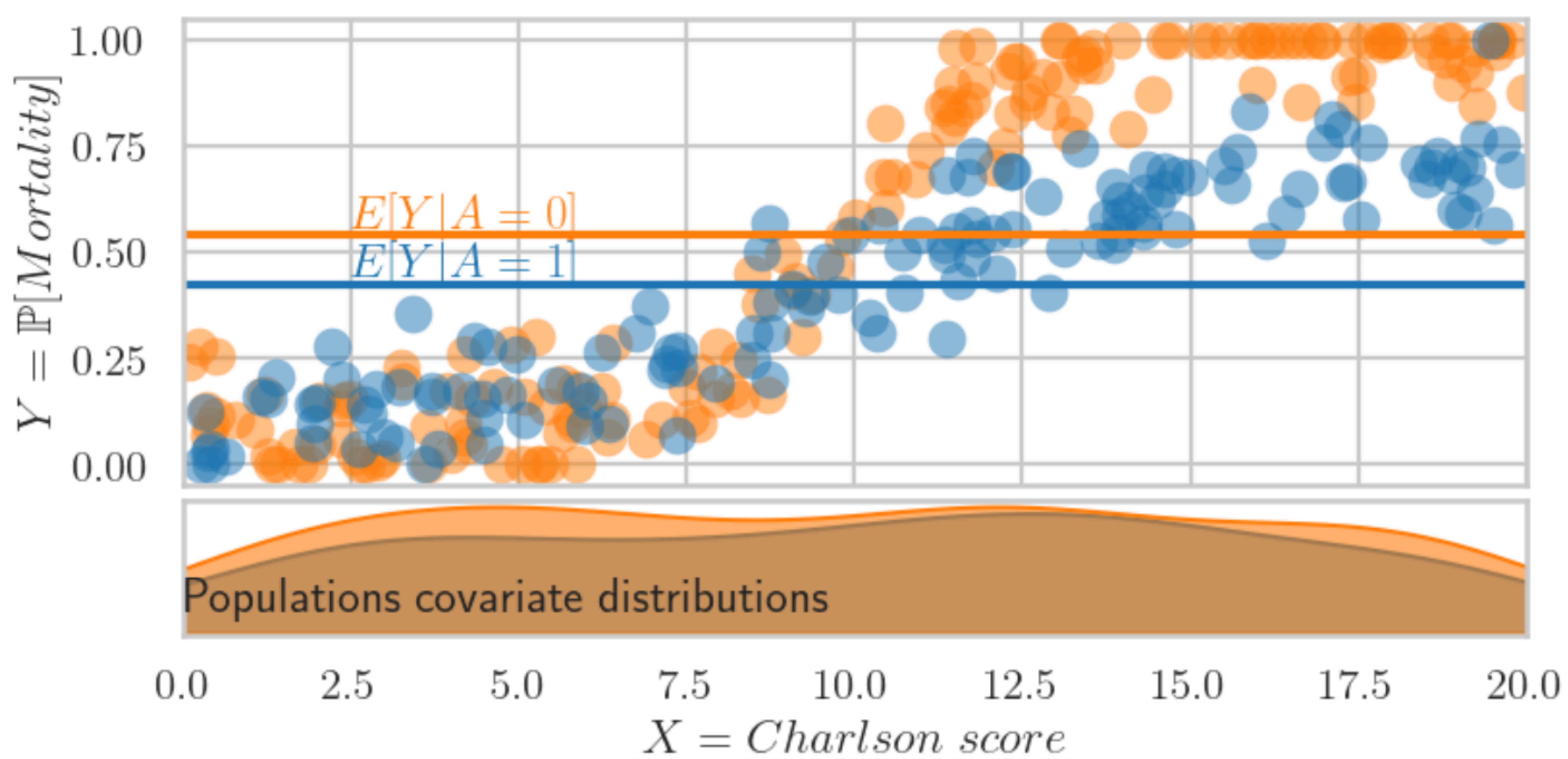
$$Y(1), Y(0) \perp\!\!\!\perp A$$

Force random assignment of the intervention

Illustration: RCT data

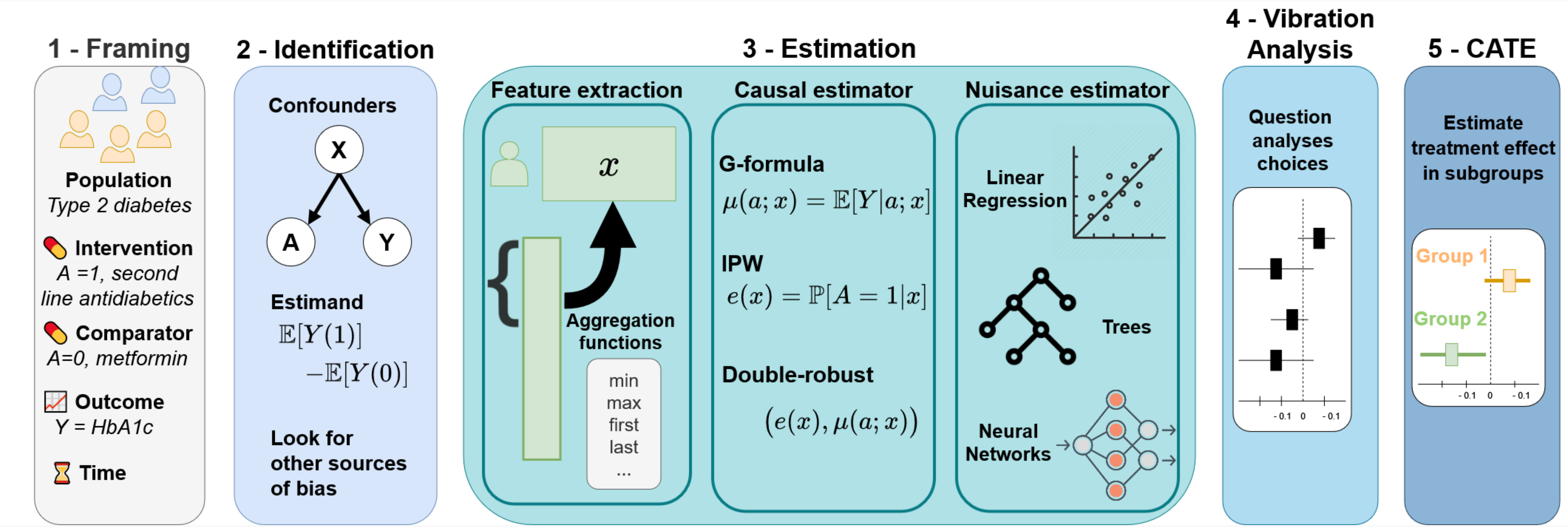


Compute the difference in mean (DM): $\tau_{\mathrm{DM}} = \mathbb{E}[Y(1)]$ - $\mathbb{E}[Y(0)]$



Four steps of causal inference : Framing, identification, statistical inference, vibration analysis

Complete inference flow



Today: Framing and identification

Framing: How to ask a sound causal question

PICO framework (Richardson et al., 1995)

Originally designed for clinical research. It is a structured approach to formulate a research question. Critical for health technology assessment (eg. Haute Autorité de santé).

PICO stands for

- Population : Who are we interested in?
- Intervention : What treatment/intervention do we study?
- Comparison : What are we comparing it to?
- Outcome : What are we interested in?

PICO framework (Richardson, Wilson, Nishikawa, & Hayward, 1995)

Originally designed for clinical research. It is a structured approach to formulate a research question. Critical for health technology assessment (eg. Haute Autorité de santé).

PICO stands for

- Population : Who are we interested in?
- Intervention : What treatment/intervention do we study?
- Comparison : What are we comparing it to?
- Outcome : What are we interested in?

Example with the job dataset (LaLonde, 1986)

Built to evaluate the impact of the National Supported Work (NSW) program. The NSW is a transitional, subsidized work experience program targeted towards people with long-standing employment problems.

The PICO framework

Component	Description	Example
Population	What is the target population of interest?	People with longstanding employment problems
Intervention	What is the intervention?	On-the-job training lasting between nine months and a year
Control	What is the relevant comparator?	No training
Outcome	What are the outcomes?	Earnings in 1978
Time	Is the start of follow-up aligned with intervention assignment?	The period of follow-up for the earning is the year after the intervention

PICO: other examples in econometrics

The Oregon Health Insurance Experiment (Finkelstein et al., 2012) : A randomized experiment by lottery assessing the impact of Medicaid on low-income adults in Oregon.

- P: Low-income adults in Oregon
- I: Medicaid
- C: No insurance
- O: Healthcare uses and expenditures, health outcomes

PICO: other examples in econometrics

The economic impact of climate change on US agricultural land. (Deschênes & Greenstone, 2007): difference-in-differences design assessing the impact of climate change on agricultural profits.

- P: US agricultural land
- I: Climate change
- C: No climate change
- O: Agricultural profits

PICO: other examples in econometrics

The impact of class size on test scores. (Angrist & Lavy, 1999): regression discontinuity design.

- P: Fourth and fifth grades school in Israel
- I: Class size increases by one unit
- C: No class size increase
- O: Test scores (math and reading)

Identification: List necessary information to answer the causal question

Build a causal model

The identification step builds a causal model to answer the research question.

Several important steps:

- Potential outcome framework : mathematical tool to reason about causality
- Directed acyclic graphs (DAG) : graphical tool to reason about causality
- Causal estimand : what is the targeted quantity?

Potential outcomes, (Neyman, 1923; Rubin, 1974)

The Neyman-Rubin model, let:

- Y be the outcome,
- A the (binary) treatment

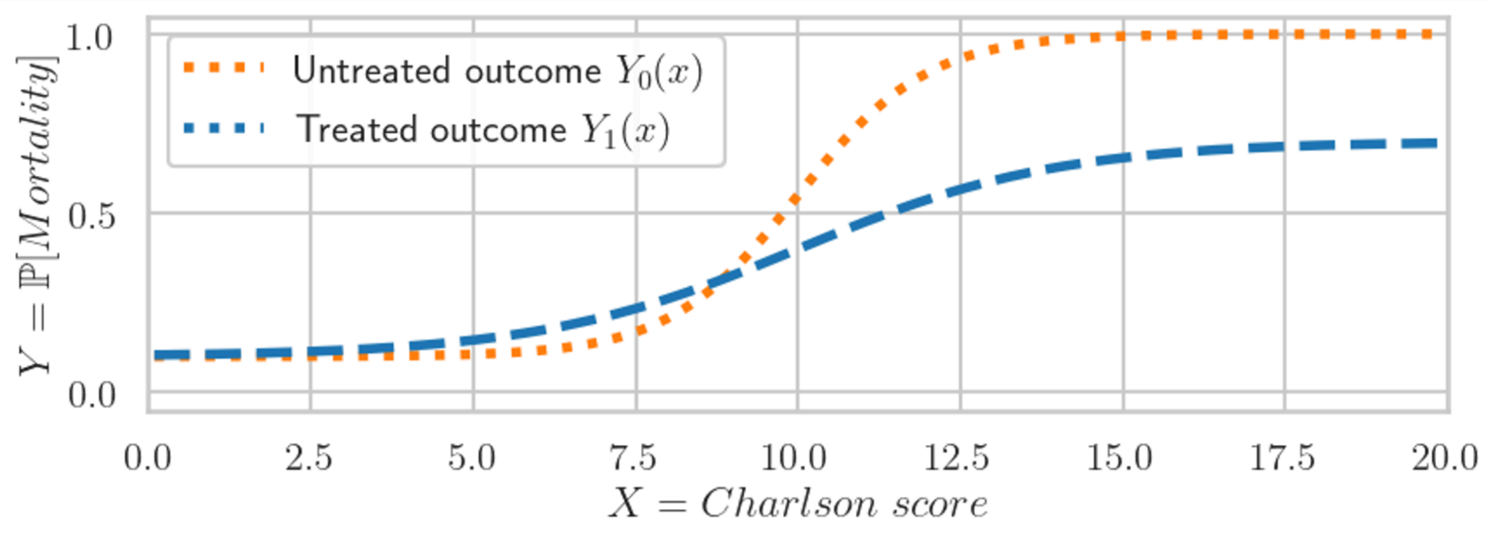
For each individual, we have two potential outcomes: $Y(1)$ and $Y(0)$. But only one is observed, depending on the treatment assignment: $Y(A)$.

Potential outcomes, (Neyman, 1923; Rubin, 1974)

The Neyman-Rubin model, let:

- Y be the outcome,
- A the (binary) treatment

For each individual, we have two potential outcomes: $Y(1)$ and $Y(0)$. But only one is observed, depending on the treatment assignment: $Y(A)$.

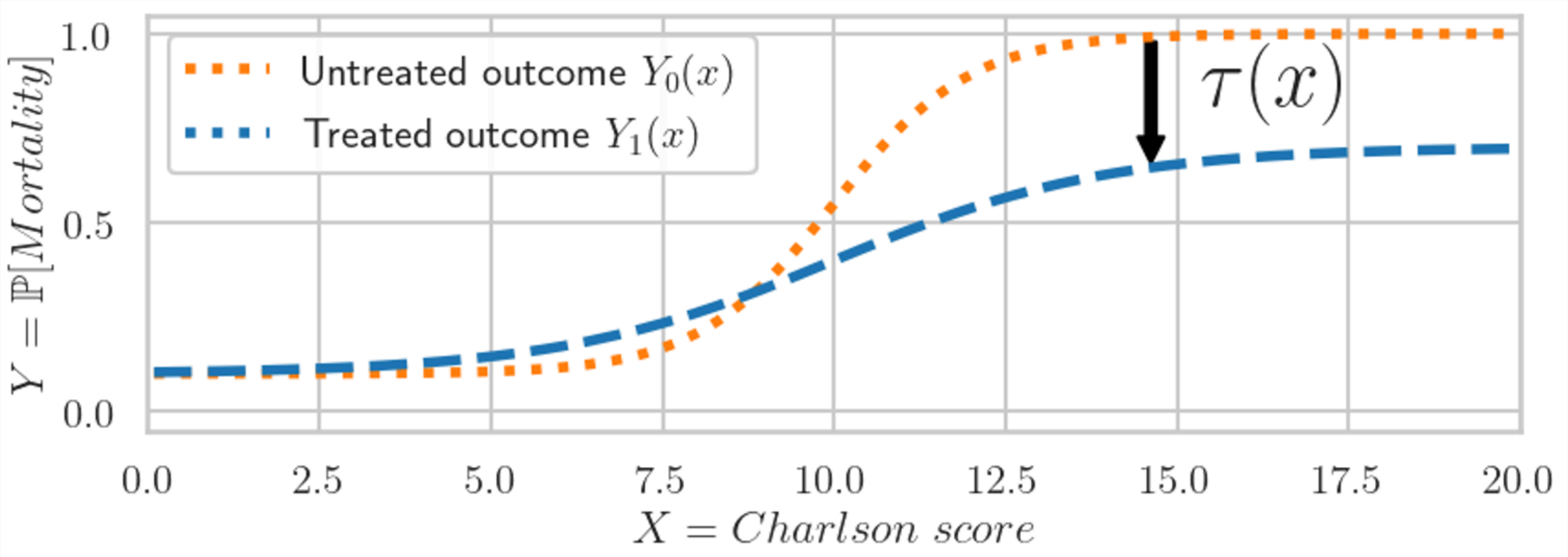


Causal estimand: What is the targeted quantity (with potential outcomes)?

- Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$
- Conditional average treatment effect (CATE): $\mathbb{E}[Y(1) - Y(0) \mid X]$

Causal estimand: What is the targeted quantity (with potential outcomes)?

- Average treatment effect (ATE): $\mathbb{E}[Y(1) - Y(0)]$
- Conditional average treatment effect (CATE): $\mathbb{E}[Y(1) - Y(0) \mid X]$



Other estimands

- Average treatment effect on the treated (ATT): $\mathbb{E}[Y(1) - Y(0) \mid A = 1]$
- Conditional average treatment effect on the treated (CATT):

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1, X]$$

Other estimands

- Average treatment effect on the treated (ATT): $\mathbb{E}[Y(1) - Y(0) \mid A = 1]$
- Conditional average treatment effect on the treated (CATT):

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1, X]$$

Other estimands more used in epidemiology

- Risk ratio (RR): $\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]}$
- Odd ratio (OR) for binary outcome: $\left(\frac{\mathbb{P}[Y(1)=1]}{\mathbb{P}[Y(1)=0]}\right) / \left(\frac{\mathbb{P}[Y(0)=1]}{\mathbb{P}[Y(0)=0]}\right)$

See (Colnet, Josse, Varoquaux, & Scornet, 2023) for a review of the different estimands and the impact on generalization.

PICO framework, link to the potential outcomes

Component	Description	Notation	Example
Population	What is the target population of interest?	$X \sim P(X)$	People with longstanding employment problems
Intervention	What is the intervention?	$A \sim P(A=1) = p_A$	On-the-job training lasting between nine months and a year
Control	What is the relevant comparator?	$1 - A \sim 1 - p_A$	No training
Outcome	What are the outcomes?	$Y(1), Y(0) \sim P(Y(1), Y(0))$	Earnings in 1978
Time	Is the start of follow-up aligned with intervention assignment?	N/A	The period of follow-up for the earning is the year after the intervention

What can we learn from the data?

- Four assumptions, referred as strong ignorability
- Required to assure identifiability of the causal estimands with observational data (Rubin, 2005)

Assumption 1: Unconfoundedness, also called ignorability

Treatment assignment is as good as random given the covariates X

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$$

- Equivalent to the conditional independence on the propensity score $e(X) \mathrel{\hat{=}} \mathbb{P}(A = 1|X)$ (Rosenbaum & Rubin, 1983):

$$\{Y(1), Y(0)\} A \mid e(X)$$

Assumption 1: Unconfoundedness, also called ignorability

Treatment assignment is as good as random given the covariates X

$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X$$

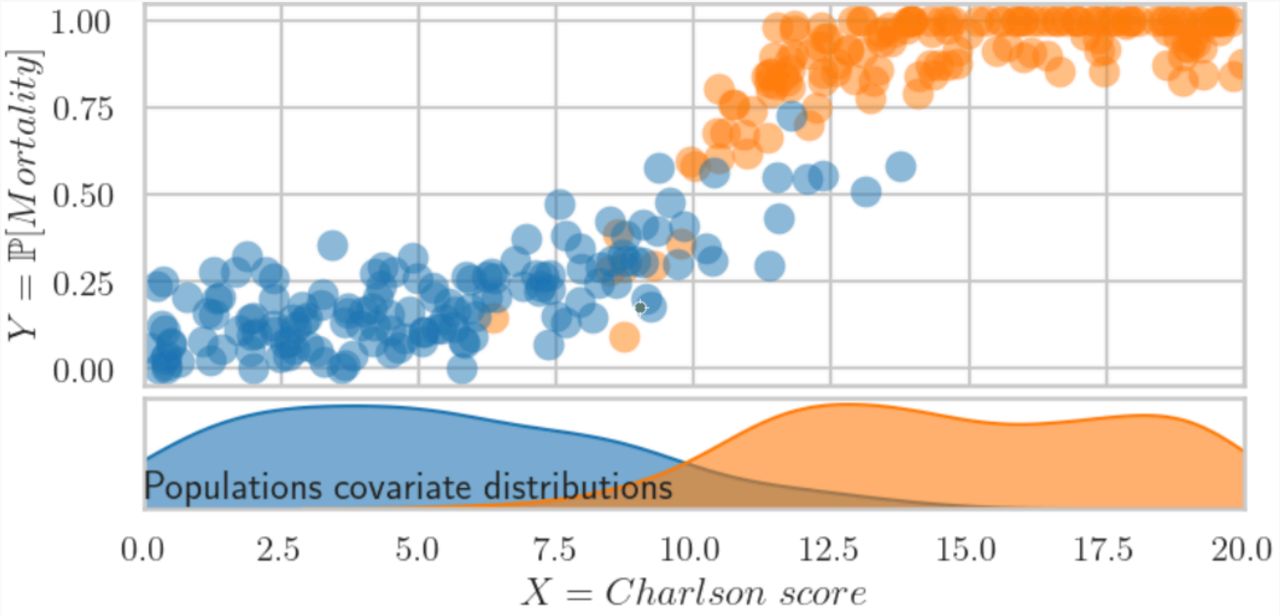
- Equivalent to the conditional independence on the propensity score $e(X) \hat{=} \mathbb{P}(A = 1|X)$ (Rosenbaum & Rubin, 1983):

$$\{Y(1), Y(0)\} A \mid e(X)$$

- Knowledge based ie. cannot be validated with data
- Because of possibly unmeasured confounders
- In practice : ask yourself if you have measured all the relevant variables that could influence both the treatment and the outcome.

The treatment is not deterministic given X

$$\eta < e(x) < 1 - \eta \text{ with } e(X) \hat{=} \mathbb{P}(A = 1|X)$$



- NB: The choices of covariates X can be viewed as a trade-off between ignorability and overlap (D'Amour et al., 2021)

Assumption 3 and 4: Consistency and generalization

Consistency, also called Stable Unit Treatment Values (SUTVA)

The observed outcome is the potential outcome of the assigned treatment for each unit i.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- The intervention A is well defined (Hernan & Robins, 2020)
- There is no interfrence ie. network effect

Assumption 3 and 4: Consistency and generalization

Consistency, also called Stable Unit Treatment Values (SUTVA)

The observed outcome is the potential outcome of the assigned treatment for each unit i.

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- The intervention A is well defined (Hernan & Robins, 2020)
- There is no interfrence ie. network effect

Generalization, also called no-covariate shift

Training and test data are are drawn from the same distribution

Directed acyclic graphs (DAG)

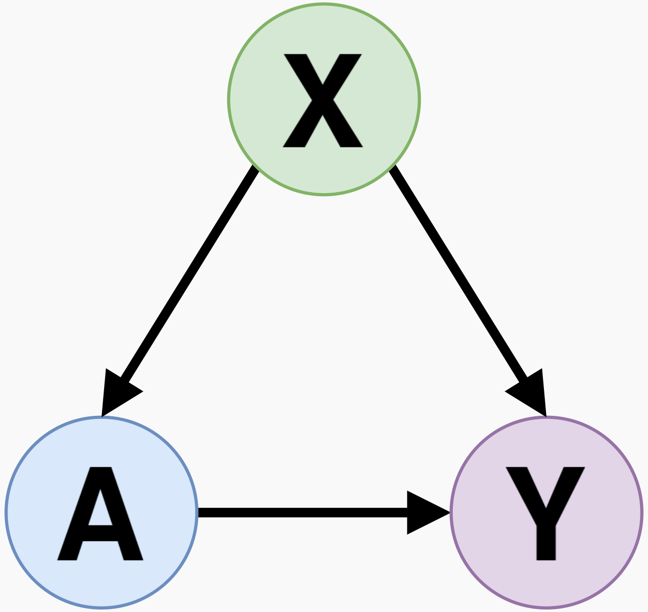
A tool to reason about causality: What are the causal status of each variable?

The confounder: a variable that influences both the treatment and the outcome.

Directed acyclic graphs (DAG)

A tool to reason about causality: What are the causal status of each variable?

The confounder: a variable that influences both the treatment and the outcome.



A word on structural equation models

Session summary

Going further

Resources

- https://web.stanford.edu/~swager/stats361.pdf
- https://www.mixtapesessions.io/
- https://alejandroschuler.github.io/mci/
- https://theeffectbook.net/index.html

Bibliography

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics, 114(2), 533–575.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ML and AI. Arxiv Preprint Arxiv:2403.02467. https://causalml-book.org/

Colnet, B., Josse, J., Varoquaux, G., & Scornet, E. (2023). Risk ratio, odds ratio, risk difference… Which causal measure is easier to generalize?. Arxiv Preprint Arxiv:2303.16008.

Deschênes, O., & Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. American Economic Review, 97(1), 354–385.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics, 221(2), 644–654.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Oregon Health Study Group, t. (2012). The Oregon health insurance experiment: evidence from the first year. The Quarterly Journal of Economics, 127(3), 1057–1106.

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. International Conference on Machine Learning, 1764–1772.

Hernan, M., & Robins, J. (2020). Causal inference: What if. boca raton: Chapman & hill/crc. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems. Arxiv Preprint Arxiv:2206.15475.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604–620.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych, 10(1), 1–51.

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. ACP Journal Club, 123(3), A12–3.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–689.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469), 322–331.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111–133.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage, 145, 166–179.

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.