Causal Inference

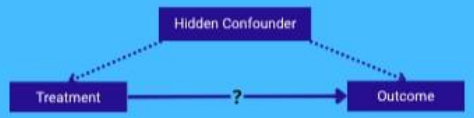
How does sample size influence Informal Benchmarking?

Introduction

Causal ML deals with **causations** rather than correlations, evaluating whether a treatment actually drives the observed outcome. However, **hidden confounders** - unmeasured variables influencing both treatment and result - can introduce bias into any study.

Sensitivity analysis evaluates the robustness of causal claims by **estimating how strong a hidden confounder** must be **to invalidate** a study's result.

- Can the outcome be explained by a latent variable rather than the treatment?



Informal Benchmarking assesses the plausibility of such a hidden confounder by **comparing the required sensitivity strength against the strength of measured variables.**

What behaviour to expect from Informal Benchmarking?

What if we have very little data? Does adding more data improve the score? At what point is my estimate stable?

Mathematical Foundations

While **informal benchmarking takes many forms**, this study focuses on estimating strength by quantifying how **the removal of specific feature subsets alters the predicted odds of treatment assignment**.

propensity score: the probability of receiving treatment given a set of features

$$p(\mathrm{t}|\mathbf{X})$$

odds of receiving treatment:

$$\Omega(\mathbf{X}) = \frac{p(\mathrm{t}|\mathbf{X})}{1 - p(\mathrm{t}|\mathbf{X})}$$

Framework: Leave-Multiple-Out (LMO) informal benchmarking algorithm.
Estimator: Logistic Regression with L2 regularization to model propensity scores.
Benchmark Estimation: the resulting **odds ratio**, where $\mathbf{S} \subset \mathbf{X}$ is any proper subset of measured features:

$$\text{Benchmark Score} = \frac{\Omega(\mathbf{X})}{\Omega(\mathbf{X} - \mathbf{S})}$$

Because of these non-linear odds ratio calculations, the framework is exceptionally **sensitive to probability estimates approaching 0 or 1**.

Methods

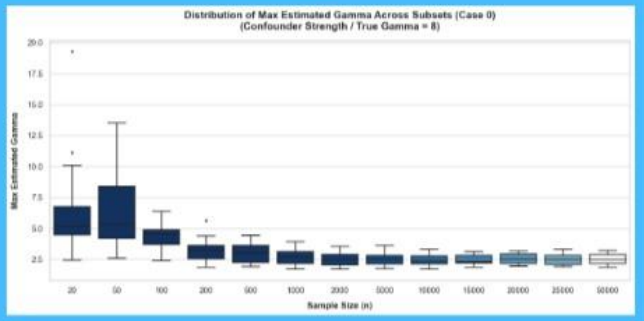
Simulation Setup:

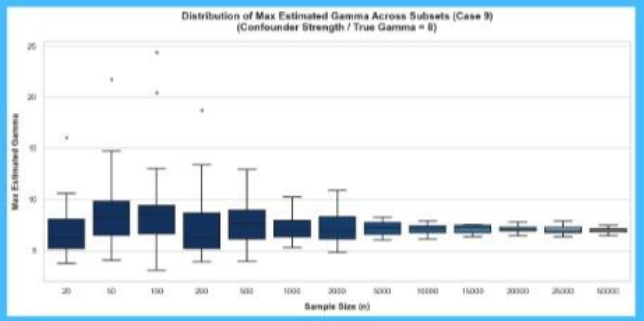
- Generated synthetic datasets spanning distinct structural scenarios
  - varying strength of measured features
  - different correlation between the hidden and measured variables
- Tested across a vast data spectrum **from N = 20 to N = 50,000**.

n = {20, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 15000, 20000, 25000, 50000}

- Monte Carlo design for empirical stability.

Experiment 1 : weak or uncorrelated features



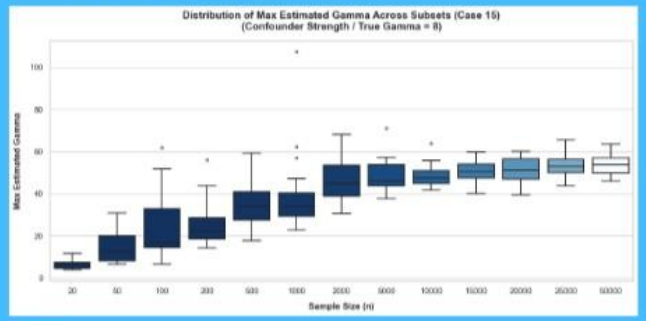


Results

Benchmark scores **flatten and stabilize as sample size increases**.

In small samples, the propensity model **overfits**, producing over-confident predictions, which **inflate the benchmark**. Large sample size acts like a natural **regularizer**, forcing the model to generalize better, making it more **conservative** and **stabilizing** the final scores.

Experiment 2 : strong or highly correlated features



Results

Benchmark scores **scale steadily upward as sample size increases**.

Massive sample sizes (N = 50,000) grant well-specified models the statistical **power to accurately map the extreme tails** of the true propensity distribution. Because odds ratios are highly **sensitive to probabilities approaching 0 or 1**, capturing these genuine tail cases inherently **drives the benchmark scores upward**.

Conclusions and a Word of Caution

In our experimental setup, **because the ground truth probabilities spanned the extreme tails**, the benchmark functioned as intended,:

assigning higher scores to models that accurately captured the true distribution and lower scores to weaker alternatives as sample size increased

However, the assumption that the true distribution always encompasses rare events and outliers **does not always hold in practice**. In scenarios where it fails, we could observe the exact opposite trend, raising a pressing question:

Does informal benchmark reward well-specified models or models that output high probabilities?

This mechanic renders the benchmark highly vulnerable to overparameterized or poorly calibrated estimators that overconfidently force predictions toward 0 or 1.

Contact Information

Author: Maja Czerwińska, m.p.czerwinska-1@student.tudelft.nl
Supervisors: Jesse Krijthe, Matej Havelka