ECON 293/MGTECON 634:
Machine Learning and Causal Inference

Stefan Wager
Stanford University

Week 2: Machine Learning for
Average Treatment Effects

Spring 2021

Last week: Machine learning crash course / introduction / review.

This week: Deploying off-the-shelf machine learning tools to support inference about average treatment effects.

Next weeks: Developing machine learning based tools for estimating treatment heterogeneity and personalization.

A central goal of machine learning is to understand **what usually happens** in a given situation, e.g.,

- Given today's weather, what's the chance tomorrow's air pollution levels will be dangerously high?

Most economists want to predict **what would happen** if we changed the system, e.g.,

- How does the answer to the above question change if we reduce the number of cars on the road?

This class is about understanding how ideas from machine learning can be rigorously **adapted or deployed** for the study of what-if questions.

Roadmap: This Week

Part 1: Introduction to **average treatment effects** and randomized trials.

Part 2: Confounding and regression adjustments. Comparison of **regression adjustments** done via OLS versus generic machine learning. (Spoiler: In a non-parametric setting, using machine learning for regression adjustments can get you consistency, but not confidence intervals.)

Part 3: The **propensity score**, and inverse-propensity weighting.

Part 4: Rigorous **confidence intervals** for the average treatment effect under non-parametric confounding. The solution deploys machine learning via the **doubly robust** estimation strategy.

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Neyman, 1923; Rubin, 1974), we posit the existence of quantities $Y_i(0)$ and $Y_i(1)$, such that $Y_i = Y_i(W_i)$.

- Potential outcomes correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).
- The **causal effect** of the treatment is $Y_i(1) - Y_i(0)$.

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Our first goal is to estimate the **average treatment effect (ATE)**

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

Of course, we only get to see $Y_i = Y_i(W_i)$. This **"missing data"** issue is a fundamental problem in causal inference.

The potential outcomes framework

The simplest way to **identify** the ATE in the potential outcomes is via a **randomized trial**:

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp W_i.$$

In a randomized trial, we can check that:

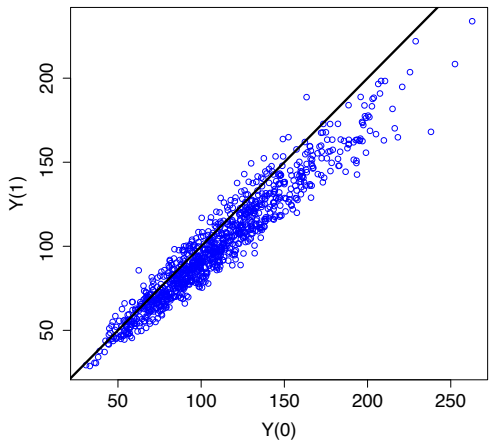
$$\begin{aligned}\tau &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\ &= \mathbb{E}\left[Y_i(1) \,\middle|\, W_i = 1\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, W_i = 0\right] \\ &= \mathbb{E}\left[Y_i \,\middle|\, W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, W_i = 0\right],\end{aligned}$$

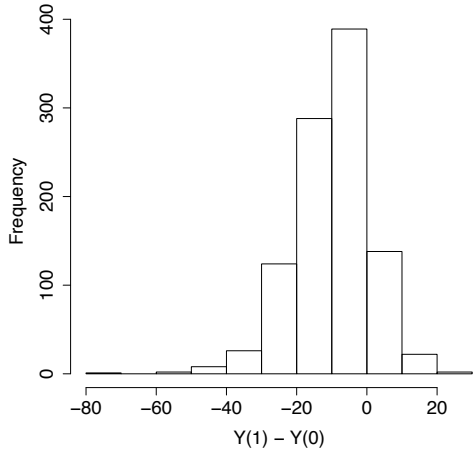
where the last line only has **observable moments**.

Thus, although we never observe $\tau_i = Y_i(1) - Y_i(0)$, we can **consistently estimate** $\tau = \mathbb{E}\left[\tau_i\right]$ in a randomized trial.

Example: The outcome Y_i is daily **air quality index**. The treatment imposes restrictions on driving to reduce traffic.

$Y_i(0)$	$Y_i(1)$	τ_i
154.68	153.49	-1.20
135.67	120.40	-15.27
103.46	117.68	14.23
117.62	95.08	-22.54
161.11	146.73	-14.39
117.89	105.05	-12.84
84.00	75.59	-8.41
73.32	65.68	-7.63
100.07	93.80	-6.28
103.81	82.30	-21.51
...	...	...
111.68	101.47	-10.21





Example: The outcome Y_i is daily **air quality index**. The treatment imposes restrictions on driving to reduce traffic.

$Y_i(0)$	$Y_i(1)$	τ_i
154.68	—	—
135.67	—	—
—	117.68	—
—	95.08	—
—	146.73	—
117.89	—	—
—	75.59	—
—	65.68	—
100.07	—	—
—	82.30	—
...	...	...
110.59	100.52	—

- In practice, we only ever observe a **single** potential outcome.
- However, in a RCT, we can use **averages** over the treated and controls to estimate the ATE.
- We **estimate** $\hat{\tau}$ as $110.59 - 100.52 = 10.07$.

Today, our main focus is on average treatment effect estimation in the presence of **confounders**.

Example:

- The outcome Y_i is quality-adjusted life years.
- The treatment W_i is about obesity mitigation and prevention.
- Covariates X_i are electronic medical records, including weight.
- Doctors are more likely to prescribe the treatment to patients with higher weight.

Why the difference in means estimator fails outside RCTs