DOTE 6635: Artificial Intelligence for Business Research

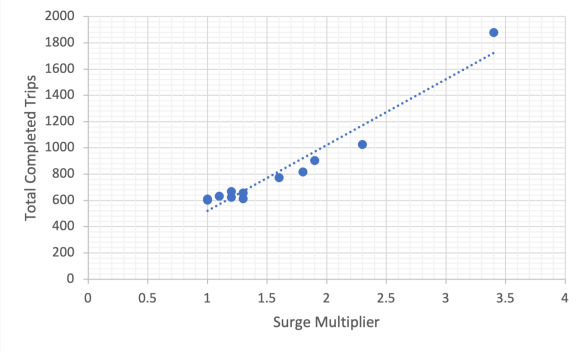
Causal Inference Fundamentals

Renyu (Philip) Zhang

1

From Philosophy to Science

- Causality has been an important philosophical question throughout human civilization (https://grok.com/share/bGVnYWN5_fabab87f-004e-4711-83bf-47e9cb4ba195).
- Neyman (1923) and Rubin (1974) make causality a scientific field of study.
- Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is part of a larger system.



[CITATION] Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes
J Neyman - Roczniki Nauk Rolniczych, 1923
☆ Save 99 Cite Cited by 527 Related articles

Mortality Rate

Treatment	Condition: Mild	Condition: Severe	Condition: Total
A	15% (210/1400)	30% (30/100)	16% (240/1500)
B	10% (5/50)	20% (100/500)	19% (105/550)

Estimating causal effects of **treatments** in **randomized** and **nonrandomized studies**.
DB Rubin - Journal of educational Psychology, 1974 - psycnet.apa.org
… use of carefully controlled **nonrandomized** data to **estimate causal effects** is a reasonable … of the use of **nonrandomized studies** to **estimate causal effects** of **treatments** (eg, Campbell & …
☆ Save 99 Cite Cited by 12906 Related articles All 12 versions

2

2

Agenda

- Potential Outcomes Model, RCT and Difference-in-Mean Estimator
- Linear Regression
- Unconfoundedness and Inverse Propensity Weighting
- Augmented Inverse Propensity Weighting and Double Robustness

3

3

Potential Outcomes Model

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- In total n subjects, each subject is assigned a binary treatment W_i = 0,1 and an outcome $Y_i(W_i)$.

The **individual causal effect** of the treatment on the i-th unit is then[2]

$$\Delta_i = Y_i(1) - Y_i(0). \qquad (1.1)$$

- Fundamental challenge of causal inference: Only one of $Y_i(1)$ and $Y_i(0)$ is observable, so we take average:

$$\overline{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right)$$

Sample Average Treatment Effect

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

Average Treatment Effect

$$\hat{\tau}_{DM} := \frac{1}{n_1}\sum_{W_i=1} Y_i - \frac{1}{n_0}\sum_{W_i=0} Y_i, \qquad n_w = |\{i : W_i = w\}|$$

Difference-in-Mean (DM) Estimator

4

4

RCT & SUTVA

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Under randomized controlled trials, we impose two additional assumptions:

$$Y_i = Y_i(W_i) \qquad (1.5) \qquad \text{(SUTVA)}$$
$$W_i \perp\!\!\!\perp \{Y_i(0),\, Y_i(1)\} \qquad \text{(random treatment assignment)}$$

- Stable Unit Treatment Value Assumption (SUTVA) = Consistency + No Interference.
- Usually, we need unbiasedness and statistical consistency first.

Suppose furthermore that the treatment is in fact randomized, i.e., that conditionally all the potential outcomes $\{Y_i(0), Y_i(1)\}_{i=1}^n$ and the number of treated units n_1, all units are treated with the same probability:[3]

$$\mathbb{P}\left[W_i = 1 \,\middle|\, \{Y_i(0), Y_i(1)\}_{i=1}^n, n_1\right] = \frac{n_1}{n}, \quad i = 1, \ldots, n. \tag{1.6}$$

Then $\hat{\tau}_{DM}$ is finite-sample unbiased for the SATE as defined in (1.2).

Theorem 1.1. *Under assumptions (1.5) and (1.6),*

$$\mathbb{E}\left[\hat{\tau}_{DM} \,\middle|\, \{Y_i(0), Y_i(1)\}_{i=1}^n, n_0 > 0, n_1 > 0\right] = \overline{\Delta}. \tag{1.7}$$

5

5

Statistical Inference with RCT

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- If the treatment assignment is randomized as a Bernoulli trial,

$$W_i \,\big|\, \{Y_i(0), Y_i(1)\} \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi), \quad 0 < \pi < 1. \tag{1.8}$$

- Then we have a central limit theorem (CLT), or root-n consistency:

Theorem 1.1

Theorem 1.2. *Under the assumptions of Theorem 1.2, suppose furthermore that the potential outcomes are drawn as* $\{Y_i(0), Y_i(1)\} \overset{iid}{\sim} P$ *from a distribution* P *with bounded second moments and that we run a Bernoulli trial as in (1.8). Then,*

$$\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \Rightarrow \mathcal{N}(0, V_{DM}), \quad V_{DM} = \frac{\text{Var}\left[Y_i(0)\right]}{1-\pi} + \frac{\text{Var}\left[Y_i(1)\right]}{\pi}. \tag{1.9}$$

Furthermore, the plug-in variance estimate

$$\widehat{V}_{DM} := \frac{n}{n_0^2} \sum_{W_i=0} \left(Y_i - \frac{1}{n_0} \sum_{W_i=0} Y_i\right)^2 + \frac{n}{n_1^2} \sum_{W_i=1} \left(Y_i - \frac{1}{n_1} \sum_{W_i=1} Y_i\right)^2 \tag{1.10}$$

is consistent, $\widehat{V}_{DM} \to_p V_{DM}$.

$$\lim_{n\to\infty} \mathbb{P}\left[\tau \in \left(\hat{\tau}_{DM} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_{DM}/n}\right)\right] = 1 - \alpha$$

6

6

RCT as the Gold Standard for Causal Inference

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- We usually say that RCTs are the gold standard for causal inference. Why?
- The estimator (DM) is very simple.
- It achieves root-n consistency, so we can have valid inference.
- It is unbiased for finite samples (Theorem 1.1 above).

7

7

Back to Our Original Causal Inference Problem

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- We have made three essential assumptions that make our problem of inferring ATE super easy.
  1. Identical and independently distributed (IID) samples
  2. Random treatment assignment (RCT)
  3. SUTVA
- Next, we try to relax them one by one.

8

8

Agenda

- Potential Outcomes Model, RCT and Difference-in-Mean Estimator
- Linear Regression
- Unconfoundedness and Inverse Propensity Weighting
- Augmented Inverse Propensity Weighting and Double Robustness

9

9

Power of Covariates

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- We now assume that there are some control variables X in the (linear) data generating process (DGP).

Linear DGP

$$Y_i(w) = \alpha_{(w)} + X_i \cdot \beta_{(w)} + \varepsilon_i(w),$$
$$\mathbb{E}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = 0, \quad \mathrm{Var}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = \sigma^2$$
$$\mathbb{P}\left[W_i = 0\right] = \mathbb{P}\left[W_i = 1\right] = \frac{1}{2} \quad \mathbb{E}\left[X\right] = 0, \quad \text{and define} \quad A = \mathrm{Var}\left[X\right]$$

Linear Regressions under Different Treatment Assignments

$$Y_i \sim \alpha_{(0)} + X_i \cdot \beta_{(0)} \text{ for all } i \text{ with } W_i = 0,$$
$$Y_i \sim \alpha_{(1)} + X_i \cdot \beta_{(1)} \text{ for all } i \text{ with } W_i = 1,$$

Estimator for ATE

$$\hat{\tau}_{IREG} = \hat{\alpha}_{(1)} - \hat{\alpha}_{(0)} + \overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right).$$

10

10

Power of Covariates

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- We now assume that the DGP is linear in some control variables X.

CLT for OLS

$$\sqrt{n_w}\left(\begin{pmatrix}\hat{\alpha}_{(w)}\\ \hat{\beta}_{(w)}\end{pmatrix}-\begin{pmatrix}\alpha_{(w)}\\ \beta_{(w)}\end{pmatrix}\right)\Rightarrow\mathcal{N}\left(0,\,\sigma^2\begin{pmatrix}1 & 0\\ 0 & A^{-1}\end{pmatrix}\right)$$

CLT for IREG

$$\hat{\tau}_{IREG}-\tau=\underbrace{\hat{\alpha}_{(1)}-\alpha_{(1)}}_{\approx\mathcal{N}(0,\sigma^2/n_1)}-\underbrace{\hat{\alpha}_{(0)}-\alpha_{(0)}}_{\approx\mathcal{N}(0,\sigma^2/n_0)}+\underbrace{\overline{X}\left(\beta_{(1)}-\beta_{(0)}\right)}_{\approx\mathcal{N}\left(0,\,\left\|\beta_{(1)}-\beta_{(0)}\right\|_A^2/n\right)}$$

$$+\underbrace{\overline{X}\left(\hat{\beta}_{(1)}-\hat{\beta}_{(0)}-\beta_{(1)}+\beta_{(0)}\right)}_{\mathcal{O}_P(1/n)},$$

$$\sqrt{n}\left(\hat{\tau}_{IREG}-\tau\right)\Rightarrow\mathcal{N}\left(0,\,V_{IREG}\right),\qquad V_{IREG}=4\sigma^2+\left\|\beta_{(0)}-\beta_{(1)}\right\|_A^2$$

11

11

Power of Covariates

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- We now assume that the DGP is linear in some control variables X.

Variance for DM

$$\begin{aligned}V_{DM}&=\frac{\text{Var}\left[Y_i(0)\right]}{0.5}+\frac{\text{Var}\left[Y_i(1)\right]}{0.5}\\&=2\left(\text{Var}\left[X_i\beta_{(0)}\right]+\sigma^2\right)+2\left(\text{Var}\left[X_i\beta_{(1)}\right]+\sigma^2\right)\\&=4\sigma^2+2\left\|\beta_{(0)}\right\|_A^2+2\left\|\beta_{(1)}\right\|_A^2\\&=4\sigma^2+\left\|\beta_{(0)}+\beta_{(1)}\right\|_A^2+\left\|\beta_{(0)}-\beta_{(1)}\right\|_A^2,\end{aligned}$$

IREG More Efficient Than DM

$$V_{IREG}=V_{DM}-\left\|\beta_{(0)}+\beta_{(1)}\right\|_A^2\leq V_{DM}$$

12

12

Non-linear DGPs

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Consider a model where the DGP is non-linear (still random treatment assignment).

Non-linear DGP

$$\mu_{(w)}(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad \sigma^2_{(w)}(x) = \mathrm{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$$

$$\mathrm{P}[W_i = 1] = \pi \qquad \mathbb{E}[X] = 0, \quad \text{and define} \quad A = \mathrm{Var}[X]$$

Variance of DM

$$\sqrt{n}\left(\hat{\tau}_{DM} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{DM}\right) = 4\sigma^2 + 2\,\mathrm{Var}\left[\mu_{(0)}(X_i)\right] + 2\,\mathrm{Var}\left[\mu_{(1)}(X_i)\right]$$

Same Estimator for ATE

$$Y_i \sim \alpha_{(0)} + X_i \cdot \beta_{(0)} \text{ for all } i \text{ with } W_i = 0,$$
$$Y_i \sim \alpha_{(1)} + X_i \cdot \beta_{(1)} \text{ for all } i \text{ with } W_i = 1,$$
$$\hat{\tau}_{IREG} = \hat{\alpha}_{(1)} - \hat{\alpha}_{(0)} + \overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right).$$

13

13

Non-linear DGPs

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Consider a model where the DGP is non-linear (still random treatment assignment).

Best Linear Projection Coefficients

$$\left(\alpha^*_{(w)},\, \beta^*_{(w)}\right) = \mathrm{argmin}_{\alpha,\,\beta}\left\{\mathbb{E}\left[\left(Y_i(w) - \alpha - X_i \cdot \beta\right)^2\right]\right\}$$

CLT for IREG

Theorem 1.3. *Under the conditions of Theorem 1.2, assume furthermore that* $\mathbb{E}[X'X]$ *is invertible. Then,*

$$\sqrt{n}\left(\hat{\tau}_{IREG} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{IREG}\right),$$
$$V_{IREG} = \mathrm{Var}\left[X_i \cdot \left(\beta^*_{(1)} - \beta^*_{(0)}\right)\right] + \frac{1}{\pi}\mathbb{E}\left[\left(Y_i(1) - \alpha^*_{(1)} - X_i \cdot \beta^*_{(1)}\right)^2\right] \quad (1.23)$$
$$+ \frac{1}{1-\pi}\mathbb{E}\left[\left(Y_i(0) - \alpha^*_{(0)} - X_i \cdot \beta^*_{(0)}\right)^2\right].$$

14

14

Where Amazing Happens

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Consider a model where the DGP is non-linear (still random treatment assignment).

IREG is more efficient than DM even for nonlinear DGP!

$$
\begin{aligned}
V_{IREG} &= 2MSE^*_{(0)} + 2MSE^*_{(1)} + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|_A^2 \\
&= 4\sigma^2 + 2\,\mathrm{Var}\left[\mu_{(0)}(X) - X\beta^*_{(0)}\right] \\
&\quad + 2\,\mathrm{Var}\left[\mu_{(1)}(X) - X\beta^*_{(1)}\right] + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|_A^2 \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(0)}\right]\right) \\
&\quad + 2\left(\mathrm{Var}\left[\mu_{(1)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(1)}\right]\right) + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|_A^2 \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) \\
&\quad + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|_A^2 - 2\left\|\beta^*_{(0)}\right\|_A^2 - 2\left\|\beta^*_{(1)}\right\|_A^2 \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|_A^2 \\
&= V_{DM} - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|_A^2 .
\end{aligned}
$$

$$
\hat{\tau}_{IREG} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\left(\hat{\alpha}_{(1)} + X_i\hat{\beta}_{(1)}\right)}_{\hat{\mu}_{(1)}(X_i)} - \underbrace{\left(\hat{\alpha}_{(0)} + X_i\hat{\beta}_{(0)}\right)}_{\hat{\mu}_{(0)}(X_i)}\right)
$$

Another Perspective:
Difference in Predictions

$$
\pi = 0.5
$$
$$
\sigma^2_{(1)}(x) = \sigma^2_{(0)}(x) = \sigma^2 \text{ for all } x
$$

15

15

RCT Recap

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Causal inference is to infer about a parallel world, so it must rely on taking averages.
- RCT is the gold standard for causal inference, thanks to 3 important assumptions.
  - IID
  - RCT
  - SUTVA
- DM estimator in RCT is root-n consistent and unbiased.
- With informative covariates, IREG estimator based on OLS is more efficient than DM estimator, even when the model is mis-specified.

16

16

What If RCT is Missing?

RCT vs. Observational Studies at FB Ads: https://pubsonline.informs.org/doi/10.1287/mksc.2018.1135

informs
http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 38, No. 2, March–April 2019, pp. 193–225
ISSN 0732-2399 (print), ISSN 1526-548X (online)

A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook

Brett R. Gordon,[a] Florian Zettelmeyer,[a,b] Neha Bhargava,[c] Dan Chapsky[c]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; [b] National Bureau of Economic Research, Cambridge, Massachusetts 02138; [c] Facebook Inc., Menlo Park, California 94025

Contact: b-gordon@kellogg.northwestern.edu, http://orcid.org/0000-0001-9081-569X (BRG); f-zettelmeyer@kellogg.northwestern.edu (FZ); nehab@fb.com (NB); chapsky@fb.com (DC)

Received: November 19, 2017
Revised: April 13, 2018; August 29, 2018
Accepted: September 8, 2018
Published Online in Articles in Advance: April 4, 2019

https://doi.org/10.1287/mksc.2018.1135

Abstract. Measuring the causal effects of digital advertising remains challenging despite the availability of granular data. Unobservable factors make exposure endogenous, and advertising's effect on outcomes tends to be small. In principle, these concerns could be addressed using randomized controlled trials (RCTs). In practice, few online ad campaigns rely on RCTs and instead use observational methods to estimate ad effects. We assess empirically whether the variation in data typically available in the advertising industry enables observational methods to recover the causal effects of online advertising. Using data from 15 U.S. advertising experiments at Facebook comprising 500 million user-experiment observations and 1.6 billion ad impressions, we contrast the experimental results to those obtained from multiple observational models. The observational methods often fail to produce the same effects as the randomized experiments, even after conditioning on extensive demographic and behavioral variables. In our setting, advances in causal inference methods do not allow us to isolate the exogenous variation needed to estimate the treatment effects. We also characterize the incremental explanatory power our data would require to enable observational methods to successfully measure advertising effects. Our findings suggest that commonly used observational approaches based on the data usually available in the industry often fail to accurately measure the true effect of advertising.

History: K. Sudhir served as the editor-in-chief and Anja Lambrecht served as associate editor for this article.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2018.1135.

Keywords: digital advertising • field experiments • causal inference • observational methods • advertising measurement

Comprehensive, individual-level, but observational data are NOT adequate to yield reliable estimates of causal effects.

Question: What could be the Savior without RCT?

A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook
BR Gordon, F Zettelmeyer, N Bhargava… - Marketing …, 2019 - pubsonline.informs.org
… experimental design of the 15 advertising RCTs we analyze: how advertising works at Facebook, how Facebook implements RCTs, and what determines advertising exposure. In …
☆ Save 99 Cite Cited by 329 Related articles All 18 versions Web of Science: 101 » 17

17

What If RCT is Missing?

LaLond (1986) after Nearly Four Decades: https://arxiv.org/abs/2406.00827

LaLonde (1986) after Nearly Four Decades: Lessons Learned

Guido Imbens Yiqing Xu

June 4, 2024

Abstract

In 1986, Robert LaLonde published an article that compared nonexperimental estimates to experimental benchmarks (LaLonde 1986). He concluded that the nonexperimental methods at the time could not systematically replicate experimental benchmarks, casting doubt on the credibility of these methods. Following LaLonde's critical assessment, there have been significant methodological advances and practical changes, including (*i*) an emphasis on estimators based on unconfoundedness, (*ii*) a focus on the importance of overlap in covariate distributions, (*iii*) the introduction of propensity score-based methods leading to doubly robust estimators, (*iv*) a greater emphasis on validation exercises to bolster research credibility, and (*v*) methods for estimating and exploiting treatment effect heterogeneity. To demonstrate the practical lessons from these advances, we reexamine the LaLonde data and the Imbens-Rubin-Sacerdote lottery data. We show that modern methods, when applied in contexts with significant covariate overlap, yield robust estimates for the adjusted differences between the treatment and control groups. However, this does not mean that these estimates are valid. To assess their credibility, validation exercises (such as placebo tests) are essential, whereas goodness of fit tests alone are inadequate. Our findings highlight the importance of closely examining the assignment process, carefully inspecting overlap, and conducting validation exercises when analyzing causal effects with nonexperimental data.

- Understanding the treatment assignment mechanism is critical.
- Unconfounded data + overlapping + modern methods yield robust, not necessarily valid, estimates.
- Validation exercises, such as placebo tests before treatment, are essential.

- Begin analyses of causal effects with an effort to understand the assignment mechanism. A clear grasp of the "design" is crucial for the credibility of the unconfoundedness assumption.
- Estimate the propensity score using a flexible method. Assess overlap by plotting the distributions of propensity scores for treated and control units. Trim the data based on the propensity score to make the groups more comparable.
- Apply modern methods, such as doubly-robust estimators, to estimate the average causal effects. Explore alternative estimands, such as the conditional average treatment effects and quantile treatment effects.
- Perform placebo tests, such as those using pretreatment outcomes, to validate unconfoundedness. Conduct sensitivity analyses to gauge the robustness of the findings.

Evaluating the econometric evaluations of training programs with experimental data
RJ LaLonde - The American economic review, 1986 - JSTOR
… that may help researchers evaluate other employment and training programs. First, … experimental results for the female participants, and are negative and smaller than the experimental …
☆ Save 99 Cite Cited by 3052 Related articles All 16 versions Web of Science: 964 » 18

18

Agenda

- Potential Outcomes Model, RCT and Difference-in-Mean Estimator
- Linear Regression
- Unconfoundedness and Inverse Propensity Weighting
- Augmented Inverse Propensity Weighting and Double Robustness

19

19

What if We Have No Randomized Assignment

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Simpson's Paradox: https://en.wikipedia.org/wiki/Simpson%27s_paradox

- Simpson's Paradox is a notorious example where RCT is violated.
- Solution: Aggregate DM estimates from different groups by sample size weights.
- Unconfoundedness: Conditional independence assumption (CIA, the most common, but also very strong, assumption in observational studies):

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i = x, \quad \text{for all} \quad x \in \mathcal{X}$$

→ You observe all potential confounders.

- Conditional Average Treatment Effect (CATE): $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$
- Stratified estimator: $\hat{\tau}_{STRAT} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x), \quad \hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x, W_i = 0\}} Y_i$

$$\text{where } n_x = |\{i : X_i = x\}| \text{ and } n_{xw} = |\{i : X_i = x,\, W_i = w\}|$$

20

20

Stratified Estimator

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Is the Stratified Estimator convergent?
  - CLT for each stratum:

$$\sqrt{n_x}\left(\hat{\tau}(x) - \tau(x)\right) \Rightarrow \mathcal{N}\left(0,\ \frac{\sigma^2_{(1)}}{e(x)} + \frac{\sigma^2_{(0)}(x)}{1-e(x)}\right)$$

$$\sigma^2_{(w)}(x) = \mathrm{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$$
$$e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$$
$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]$$

Propensity Score (PS)

Stratified Estimator is root-n consistent:

Theorem 2.1. *Suppose that* $\{X_i, Y_i(0), Y_i(1), W_i\} \overset{iid}{\sim} P$ *for some distribution* P *where* X_i *takes values in a finite cardinality set* $\mathcal{X}$ *and potential outcomes have bounded second moments conditionally on* X_i*. Suppose furthermore that both* (2.1) *and SUTVA hold, and that there is non-trivial treatment variation for each* $x \in \mathcal{X}$*, i.e., writing* $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$*, we have* $0 < e(x) < 1$ *for all* x*. Then, using notation as in* (1.21)*,*

(2.1): Unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\middle|\, X_i = x, \quad \text{for all} \quad x \in \mathcal{X}$$

$$\sqrt{n}\left(\hat{\tau}_{STRAT} - \tau\right) \Rightarrow \mathcal{N}\left(0,\ V_{STRAT}\right)$$
$$V_{STRAT} = \mathrm{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma^2_{(1)}(X_i)}{e(X_i)} + \frac{\sigma^2_{(0)}(X_i)}{1-e(X_i)}\right]. \qquad (2.4)$$

21

21

Stratified Estimator & Propensity Score

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Stratified estimator is just propensity score estimation on a discrete set.
  - What if the covariates are continuous?

- Unconfoundedness: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\middle|\, X_i, \qquad (2.5)$

Statistically, a key property of the propensity score is that it is a balancing score: If (2.5) holds, then in fact

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\middle|\, e(X_i), \qquad (2.7)$$

i.e., it actually suffices to control for $e(X)$ rather than X to remove biases associated with a non-random treatment assignment. We can verify this claim as follows:

$$\begin{aligned}
&\mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(0), Y_i(1)\}, e(X_i)\right] \\
&= \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(w)\}, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, \{Y_i(w)\}, e(X_i)\right]\, dx \\
&= \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, \{Y_i(w)\}, e(X_i)\right]\, dx \qquad \text{(unconf.)} \\
&= \begin{cases} e(X_i) & \text{if w = 1,} \\ 1 - e(X_i) & \text{else.} \end{cases}
\end{aligned}$$

- Overlapping assumption: $\eta \leq e(x) \leq 1 - \eta$ for all $x \in \mathcal{X}$

22

22

Propensity Stratification

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Imbens, G.W. and Rubin, D.B., 2015. *Causal inference in statistics, social, and biomedical sciences.*
Rosenbaum, P.R. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), pp.41-55.

Propensity stratification One instantiation of this idea is propensity stratification, which proceeds as follows. First obtain an estimate $\hat{e}(x)$ of the propensity score via non-parametric regression, and choose a number of strata J. Then:

1. Sort the observations according to their propensity scores, such that

$$\hat{e}\left(X_{i_1}\right) \leq \hat{e}\left(X_{i_2}\right) \leq \ldots \leq \hat{e}\left(X_{i_n}\right). \tag{2.8}$$

2. Split the sample into J evenly size strata using the sorted propensity score and, in each stratum $j = 1, ..., J$, compute the simple difference-in-means treatment effect estimator for the stratum:

$$\hat{\tau}_j = \frac{\sum_{j=\lfloor (j-1)n/J \rfloor +1}^{\lfloor jn/J \rfloor} W_i Y_i}{\sum_{j=\lfloor (j-1)n/J \rfloor +1}^{\lfloor jn/J \rfloor} W_i} - \frac{\sum_{j=\lfloor (j-1)n/J \rfloor +1}^{\lfloor jn/J \rfloor} (1-W_i) Y_i}{\sum_{j=\lfloor (j-1)n/J \rfloor +1}^{\lfloor jn/J \rfloor} (1-W_i)}. \tag{2.9}$$

3. Estimate the average treatment by applying the idea of (2.3) across strata:

$$\hat{\tau}_{PSTRAT} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_j. \tag{2.10}$$

Propensity Stratification is consistent when the propensity score estimator is uniformly consistent and the number of strata J grows appropriately with n.

Another more popular approach: Propensity Score Matching (PSM)
- Chapter 15 of Imbens and Rubin (2015)

See Rosenbaum and Rubin (1983) for comprehensive discussions on these two methods which reply on similar assumptions.

Inverse Propensity Weighting (IPW)

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Another common way of using propensity score is the IPW estimator:

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i) Y_i}{1-\hat{e}(X_i)} \right)$$

- To analyze this estimator, it is convenient to compare it to an oracle where the actual propensity score is known, because these two estimators' difference can be bounded by the performance of PS estimator:

$$\hat{\tau}^*_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i) Y_i}{1-e(X_i)} \right)$$

- A related concept, Horvitz-Thompson estimator: https://grok.com/share/bGVnYWN5_2e4d6004-5e65-4cb4-82d8-8c5aa6bc3218

Performance of IPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

Theorem 2.2. *Suppose that* $\{X_i, Y_i(0), Y_i(1), W_i\} \overset{iid}{\sim} P$, *that both* (2.5) *and SUTVA hold, and that all moments used in the expression for* V_{IPW^*} *below are finite. Then, the oracle IPW estimator is unbiased,* $\mathbb{E}\left[\hat{\tau}^*_{IPW}\right] = \tau$, *and*

$$\sqrt{n}\left(\hat{\tau}^*_{IPW} - \tau\right) \Rightarrow \mathcal{N}\left(0, V_{IPW^*}\right)$$

$$V_{IPW^*} = \mathrm{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\left(\mu_{(0)}(X_i) + (1 - e(X_i))\tau(X_i)\right)^2}{e(X_i)(1 - e(X_i))}\right] + \mathbb{E}\left[\frac{\sigma^2_{(1)}(X_i)}{e(X_i)} + \frac{\sigma^2_{(0)}(X_i)}{1 - e(X_i)}\right]. \quad (2.13)$$

Unconfoundedness:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i, \quad (2.5)$$

Overlapping:

$$\eta \leq e(x) \leq 1 - \eta \text{ for all } x \in \mathcal{X}$$

25

25

IPW vs. STRAT

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

- Stratified estimator is a special case of IPW:

$$\hat{\tau}_{STRAT} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i)\, Y_i}{1 - \hat{e}(X_i)}\right), \quad \hat{e}(x) = \frac{n_{x1}}{n_x}$$

- What is interesting is that STRAT is more efficient than IPW:

$$V_{IPW^*} = V_{STRAT} + \mathbb{E}\left[\frac{\left(\mu_{(0)}(X_i) + (1 - e(X_i))\tau(X_i)\right)^2}{e(X_i)(1 - e(X_i))}\right]$$

- When X is discrete with a natural but specific propensity model, one feasible IPW can outperform the oracle IPW.
  - This result should not be over-generalized.

26

26

What's Wrong with IPW?

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- If the propensity score e(.) is unknow, does IPW converge sufficiently fast to ensure semi-parametric efficiency, i.e., root-n consistency or the gap between oracle IPW and general IPW is of order o(sqrt(n))?

$$\hat{\tau}_{IPW} = \underbrace{\hat{\tau}^*_{IPW}}_{\text{a good estimator}} + \underbrace{\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW}}_{\text{due to errors in } \hat{e}(\cdot)}.$$

- Not good enough!
- For most ML algorithms, RMSE >> sqrt(1/n).
- So the second term is non-negligible in finite samples.
- This naïve plug-in IPW will generally produce invalid confidence intervals.

Let's try to **bound the error** using Cauchy-Schwarz:

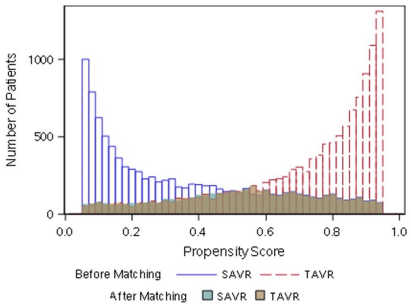
$$\begin{aligned}
&\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{W_i}{\hat{e}(X_i)} - \frac{(1-W_i)}{1-\hat{e}(X_i)}\right) - \left(\frac{W_i}{e(X_i)} - \frac{(1-W_i)}{1-e(X_i)}\right)\right) Y_i \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{W_i}{\hat{e}(X_i)} - \frac{(1-W_i)}{1-\hat{e}(X_i)}\right) - \left(\frac{W_i}{e(X_i)} - \frac{(1-W_i)}{1-e(X_i)}\right)\right)^2} \\
&\quad \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} \\
&\asymp \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{e}(X_i) - e(X_i))^2}.
\end{aligned}$$

27

Overlapping Assumption $\eta \le e(x) \le 1-\eta$ for all $x \in \mathcal{X}$

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Imbens, G.W. and Rubin, D.B., 2015. *Causal inference in statistics, social, and biomedical sciences.*
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- Overlapping assumption is important, both theoretically and practically, but we often do not have overlapping.
  - Severe weights in extreme cases, which lead to biases in estimates.
  - Estimating ATE becomes impossible, but only the ATE on overlapping samples (i.e., conditional ATE).
- Adjustments methods
  - Regression: Model sensitive to extreme data
  - Matching: Many samples are excluded
  - Stratification: Adds bias due to residual imbalance within strata
- Trimming
  - Remove data samples with PS outside [a, 1-a], a=0.1 as a baseline.
  - Further remove data samples whose PS is below q-quantile of treated units.
  - Further remove data samples whose PS is above (1-q)-quantile of control units.
  - Winsorize all PS below a to a and above 1-a to 1-a.



28

Balancing Weights

Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html
Balancing covariates via propensity score weighting: https://arxiv.org/abs/1404.1785

- IPW is a special case of balancing weights.
  - Assume density of the observed covariates, $f(x)$, exists wrt a measure μ
  - Consider a target population, denoted by a density $g(x)$, possibly different from $f(x)$
  - The ratio $h(x) = g(x)/f(x)$ is called a *tilting function*, which re-weights the observed sample to represent the target population
  - A new class of estimands: the ATE over the target population g

$$\tau_h \equiv \mathbb{E}_g[Y_i(1) - Y_i(0)] = \frac{\int \tau(x) f(x) h(x) \mu(dx)}{\int f(x) h(x) \mu(dx)} = \frac{\mathbb{E}\{h(x)\tau(x)\}}{\mathbb{E}\{h(x)\}}$$

29

29

Balancing Weights

Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html
Balancing covariates via propensity score weighting: https://arxiv.org/abs/1404.1785

- Different weighting schemes:

target population	$h(x)$	estimand	weight (w_1, w_0)
combined	1	ATE	$\left(\frac{1}{e(x)}, \frac{1}{1-e(x)}\right)$ [HT]
treated	$e(x)$	ATT	$\left(1, \frac{e(x)}{1-e(x)}\right)$
control	$1-e(x)$	ATC	$\left(\frac{1-e(x)}{e(x)}, 1\right)$
overlap	$e(x)(1-e(x))$	ATO	$(1-e(x), e(x))$
truncated combined	$\mathbf{1}(\alpha < e(x) < 1-\alpha)$		$\left(\frac{\mathbf{1}(\alpha<e(x)<1-\alpha)}{e(x)}, \frac{\mathbf{1}(\alpha<e(x)<1-\alpha)}{1-e(x)}\right)$
matching	$\min\{e(x), 1-e(x)\}$		$\left(\frac{\min\{e(x),1-e(x)\}}{e(x)}, \frac{\min\{e(x),1-e(x)\}}{1-e(x)}\right)$

Table 1: Examples of balancing weights and corresponding target population and estimand under different h.

$$\hat{\tau}_h = \frac{\sum_i w_1(x_i) Z_i Y_i}{\sum_i w_1(x_i) Z_i} - \frac{\sum_i w_0(x_i)(1-Z_i) Y_i}{\sum_i w_0(x_i)(1-Z_i)}$$

Theorem 1. *$\hat{\tau}_h$ is a consistent estimator of τ_h.*

30

30

Agenda

- Potential Outcomes Model, RCT and Difference-in-Mean Estimator
- Linear Regression
- Unconfoundedness and Inverse Propensity Weighting
- Augmented Inverse Propensity Weighting and Double Robustness

31

31

Augmented Inverse Propensity Weighting (AIPW)

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- The performance of IPW estimators depend on:
  - The performance of PS estimator
  - The weighting mechanisms (regression, matching, stratification, etc.)
- Can we improve IPW with machine learning?
- Yes, with new doubly robust estimators: Augmented Inverse Propensity Weighting.
- Two consistent characterizations of ATE: IPW and nonparametric regression.

Nonparametric Regression

$$\begin{aligned}\tau(x) &:= \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right] \\ &= \mathbb{E}\left[Y_i(1) \,\middle|\, X_i = x\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i = x\right] \\ &= \mathbb{E}\left[Y_i(1) \,\middle|\, X_i = x, W_i = 1\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i = x, W_i = 0\right] \quad \text{(unconf)} \\ &= \mathbb{E}\left[Y_i \,\middle|\, X_i = x, W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, X_i = x, W_i = 0\right] \quad \text{(SUTVA)} \\ &= \mu_{(1)}(x) - \mu_{(0)}(x),\end{aligned}$$

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

$$\hat{\tau}_{REG} = n^{-1}\sum_{i=1}^{n}(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i))$$

IPW

$$\tau = \mathbb{E}\left[\hat{\tau}^*_{IPW}\right], \quad \hat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)\,Y_i}{1-e(X_i)}\right)$$

Estimation of **regression coefficients** when **some regressors** are **not always observed**
JM Robins, A Rotnitzky, LP Zhao - Journal of the American ..., 1994 - Taylor & Francis
... each previous **estimator** is asymptotically equivalent to **some**, usually inefficient, **estimator** in our ... that every regular asymptotic linear **estimator** of a0 is asymptotically equivalent to **some** ...
☆ Save 99 Cite Cited by 3716 Related articles All 10 versions »

32

32

Double Robustness of AIPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1-\hat{e}(X_i)}\right)$$

- Weak double robustness (DR): If either $\hat{\mu}_{(w)}(x) \approx \mu_{(w)}(x)$ or $\hat{e}(x) \approx e(x)$, AIPW is consistent.

$$\hat{\tau}_{AIPW} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)}_{\text{the regression estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1-W_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right)}_{\approx \text{ mean-zero noise}}$$

$$\hat{\tau}_{AIPW} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)}_{\text{the IPW estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i)\left(1-\frac{W_i}{\hat{e}(X_i)}\right) - \hat{\mu}_{(0)}(X_i)\left(1-\frac{1-W_i}{1-\hat{e}(X_i)}\right)\right)}_{\approx \text{ mean-zero noise}}$$

- Strong double robustness: root-n consistency and CLT.

33

33

Strong Double Robustness of AIPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- More interesting and useful strong double robustness: If the black-box predictors of mean and propensity score are reasonably good (i.e., converge to the ground-truth reasonably fast), AIPW is root-n consistent.

consistency statement given above. At a high level, strong double robustness is a claim that results of the following type exist: If we use estimators $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ that are both consistent with root-mean squared error (RMSE) decaying faster than $n^{-\alpha_\mu}$ and $n^{-\alpha_e}$ respectively, and if furthermore $\alpha_\mu + \alpha_e \geq 1/2$, then

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0, V_{AIPW}\right),$$
$$V_{AIPW} = \text{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma_0^2(X_i)}{1-e(X_i)}\right] + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right]. \quad (3.5)$$

The reason this meta-result holds is that, in general, if the RMSE of $\hat{\mu}_{(w)}(x)$ decays faster than $n^{-\alpha_\mu}$ and the RMSE of $\hat{e}(x)$ decays faster than $n^{-\alpha_e}$, then the bias of AIPW decays faster than $n^{-(\alpha_\mu+\alpha_e)}$; and, in particular, if $\alpha_\mu + \alpha_e \geq 1/2$ then the bias is lower-order on the $1/\sqrt{n}$-scale. What's remarkable about this result is that, under the same conditions, the bias of the regression estimator would in general only be bounded to order $n^{-\alpha_\mu}$ and that of IPW to order $n^{-\alpha_e}$; and so the AIPW construction succeeds in making bias substantially smaller than what either the regression or IPW estimators could achieve on their own.[17]

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}(X) - \mu_{(w)}(X)\right)^2\right]^{\frac{1}{2}}, \mathbb{E}\left[\left(\hat{e}(X) - e(X)\right)^2\right]^{\frac{1}{2}} \ll \frac{1}{\sqrt[4]{n}}$$

Can be generalized to order $n^{-\alpha_\mu}$ and $n^{-\alpha_e}$

Then the error of AIPW will be of order $n^{-(\alpha_\mu+\alpha_e)}$

As long as $\alpha_\mu + \alpha_e \geq 1/2$, the error of AIPW is smaller than $o(n^{-1/2})$, establishing the desired root-n consistency.

This is the core idea of double machine learning (DML).

34

34

AIPW vs. IPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- If we need naïve plug-in IPW to be root-n consistent and produce invalid confidence intervals, the underlying ML PS estimates should have $o(n^{-1/2})$ L2-norm error (sometimes called parametric rate), which is too optimistic in most real-world applications.
- For AIPW, it suffices for conditional mean and PS estimates to achieve $o(n^{-1/4})$ L2-norm error (sometimes called semiparametric efficient), which is realistic and pretty accurate for complex ML methods.
- AIPW translates natural and realistic assumptions on ML estimates accuracy into valid inference.
- For AIPW, we only need one moment condition to be true, and the convergency speed of the final estimator is the product of convergence speed of both estimators.

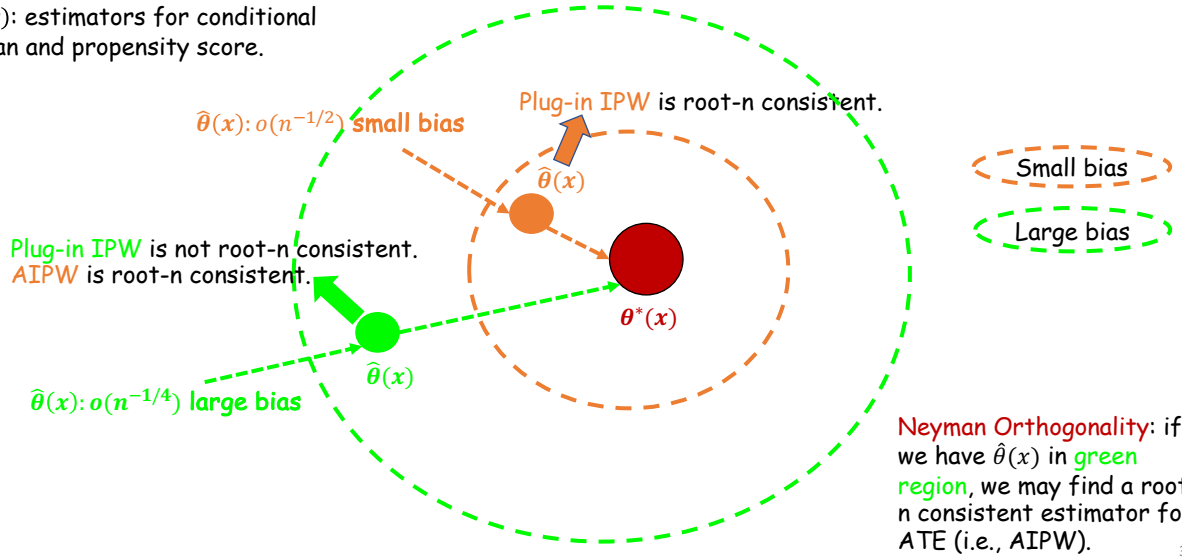
35

35

AIPW vs. IPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

$\hat{\boldsymbol{\theta}}(\boldsymbol{x})$: estimators for conditional mean and propensity score.



Neyman Orthogonality: if we have $\hat{\theta}(x)$ in green region, we may find a root-n consistent estimator for ATE (i.e., AIPW).

36

36

AIPW Asymptotic Normality

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- Like IPW, we first establish the root-n consistency of oracle AIPW, who has access to the ground-truth conditional mean and propensity score.

$$\hat{\tau}^*_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i$$

$$\Gamma_i = \mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1-W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1-e(X_i)}$$

Proposition 3.1. *Under the basic setting with SUTVA, unconfoundedness and strong overlap given at the beginning of this chapter, the oracle AIPW estimator has the limit distribution given in (3.5), i.e.,*

$$\sqrt{n}\left(\hat{\tau}^*_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0, V_{AIPW}\right). \tag{3.7}$$

$$V_{AIPW} = \mathrm{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma_0^2(X_i)}{1-e(X_i)}\right] + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right]. \tag{3.5}$$

- Hence, it suffices to show that if $\hat{\mu}_{(w)}(\cdot)$ and $\hat{e}(\cdot)$ converge to the ground-truth fast enough:

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \hat{\tau}^*_{AIPW}\right) \to_p 0$$

- To obtain this root-n equivalence between AIPW and oracle AIPW, we leverage the idea of cross-fitting.

37

Cross Fitting

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- For AIPW, we typically have two steps of estimation:
  - First, we predict the mean responses and propensity scores using ML.
  - Then, we use these predictions to estimate the ATE via AIPW.
- Cross-fitting is a common strategy for AIPW:
  - Split the data into two samples: One sample for ML prediction and the other sample to use ML predictions and data to construct AIPW estimators.

Train	Infer
Infer	Train

$$\hat{\tau}_{AIPW} = \frac{|\mathcal{I}_1|}{n}\hat{\tau}^{\mathcal{I}_1} + \frac{|\mathcal{I}_2|}{n}\hat{\tau}^{\mathcal{I}_2}, \quad \hat{\tau}^{\mathcal{I}_1} = \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i)}\right), \tag{3.9}$$

where the $\hat{\mu}^{\mathcal{I}_2}_{(w)}(\cdot)$ and $\hat{e}^{\mathcal{I}_2}(\cdot)$ are estimates of $\mu_{(w)}(\cdot)$ and $e(\cdot)$ obtained using only the half-sample $\mathcal{I}_2$, and $\hat{\tau}^{\mathcal{I}_2}$ is defined analogously (with the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ swapped). In other words, $\hat{\tau}^{\mathcal{I}_1}$ is a treatment effect estimator on $\mathcal{I}_1$ that uses $\mathcal{I}_2$ to estimate its nuisance components, and vice-versa.

- The "honest" regression residual (different training and inference data) cannot be artificially shrunk by overfitting.

38

Back to AIPW Asymptotic Normality

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- Under cross-fitting and $\hat{\mu}_{(w)}(\cdot)$ and $\hat{e}(\cdot)$ converging to ground-truth sufficiently fast, we have the root-n equivalence between between AIPW and oracle AIPW.

Theorem 3.2. *Given our basic setting with SUTVA, unconfoundedness and strong overlap, suppose that we construct $\hat{\tau}_{AIPW}$ using cross-fitting with estimators satisfying, for $w \in \{0, 1\}$ and also with the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ swapped,*

$$\text{o}(n^{-\alpha_\mu})\text{ RMSE/L}^2\text{-norm error} \rightarrow n^{-2\alpha_\mu} \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(w)}^{\mathcal{I}_2}(X_i) - \mu_{(w)}(X_i) \right)^2 \rightarrow_p 0,$$

$$\text{o}(n^{-\alpha_e})\text{ RMSE/L}^2\text{-norm error} \rightarrow n^{-2\alpha_e} \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right)^2 \rightarrow_p 0, \tag{3.11}$$

for some constants with $\alpha_\mu, \alpha_e \geq 0$ and $\alpha_\mu + \alpha_e \geq 1/2$. Then (3.9) and thus also (3.5) hold.

$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0, V_{AIPW}\right), \qquad \sqrt{n}\left(\hat{\tau}_{AIPW} - \hat{\tau}^*_{AIPW}\right) \rightarrow_p 0$$

$$V_{AIPW} = \text{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma_0^2(X_i)}{1 - e(X_i)}\right] + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right]$$

39

39

Estimation and Inference with AIPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- Generalize into K-fold cross-fitting:

We define the data into K folds (above, $K = 2$), and compute estimators $\hat{\mu}_{(w)}^{(-k)}(x)$, etc., excluding the k-th fold. Then, writing $k(i)$ as the mapping that takes an observation and puts it into one of the k folds, we can write

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{e}^{(-k(i))}(X_i)}\right). \tag{3.16}$$

Train	Train	Infer
Train	Infer	Train
Infer	Train	Train

K = 3

$$\widehat{V}_{AIPW} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\widehat{\Gamma}_i - \hat{\tau}_{AIPW}\right)^2, \rightarrow_p V_{AIPW}$$

$$\widehat{\Gamma}_i = \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{e}^{(-k(i))}(X_i)}.$$

$(1-\alpha)$-Confidence Level:

$$\tau \in \left(\hat{\tau}_{AIPW} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{1}{\sqrt{n}}\sqrt{\widehat{V}_{AIPW}}\right)$$

Inverse CDF of standard normal.

40

40

(Optimal) Efficiency of AIPW

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
Duke STA640 Causal Inference: https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html

- AIPW achieves the same efficiency as the stratify-by-X method, which is actually optimal.

$$\hat{\tau}_{STRAT} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x), \quad \hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x, W_i = 0\}} Y_i \text{ where } n_x = |\{i : X_i = x\}| \text{ and } n_{xw} = |\{i : X_i = x, W_i = w\}|$$

Theorem 3.4. *Under basic setting with SUTVA, unconfoundedness and strong overlap, V^* is the efficient variance for estimating the average treatment effect.*

$$\sqrt{n}\,(\hat{\tau} - \tau^*) \Rightarrow \mathcal{N}\,(0,\, V^*)$$

$$V^* = \text{Var}\,[\tau(X_i)] + \mathbb{E}\left[\frac{\sigma_0^2(X_i)}{1 - e(X_i)}\right] + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right]$$

- For a long time, (optimal) efficiency is not a good criterion for designing practical estimators for ATE.
  - Efficient estimators were considered fragile, complicated and/or impractical.
  - Econometric practice largely focused on methods under parametric assumptions (e.g., linear regression), or non-efficiently but conceptually simple (e.g., matching).
- The AIPW method (more generally, DML), however, makes efficient treatment effect estimators practical. See also: https://rphilipzhang.github.io/rphilipzhang/DeDL-nonblind.pdf.

41

41

Doubly Robust Policy Evaluation

Doubly robust policy evaluation and optimization
M Dudik, D Erhan, J Langford, L Li - 2014 - projecteuclid.org
... We apply the **doubly robust** technique to **policy** evaluation and **optimization** in a contextual ... For example, analogous to the results introduced in this paper, **doubly robust estimators** ...
☆ Save 99 Cite Cited by 487 Related articles All 13 versions Web of Science: 139

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf

- AIPW-like DR methods have been widely adopted in the business literature for policy evaluation.

Policy Evaluation (PE) Problem

Distribution of feature x — Randomized Policy $\nu(.|.)$

$$\nu(x, a, r) = D(x)\nu(a|x)D(r|a, x)$$

and hence $V = \mathbf{E}_\nu[r]$.

Expected Reward of Policy $\nu(.|.)$

Reward distribution given feature x and action a

1. Direct Method (DM)

$$\hat{V}_{\text{DM}} = \frac{1}{n}\sum_{k=1}^{n}\sum_{a \in \mathcal{A}} \nu(a|x_k)\hat{r}(x_k, a)$$

Estimated Reward

2. Inverse Propensity Score (IPS)

$$\hat{V}_{\text{IPS}} = \frac{1}{n}\sum_{k=1}^{n} \frac{\nu(a_k|x_k)}{\hat{\mu}_k(a_k|x_k)} \cdot r_k$$

Estimated policy from which the data is sampled.

4. Doubly Robust (DR)

$$(3.1)\quad \hat{V}_{\text{DR}} = \frac{1}{n}\sum_{k=1}^{n}\left[\hat{r}(x_k, \nu) + \frac{\nu(a_k|x_k)}{\hat{\mu}_k(a_k|x_k)} \cdot (r_k - \hat{r}(x_k, a_k))\right],$$

where

$$\hat{r}(x, \nu) = \sum_{a \in \mathcal{A}} \nu(a|x)\hat{r}(x, a)$$

42

42

DR for RL Policy Evaluation

Dynamic **coupon targeting** using batch **deep reinforcement learning**: An application to **livestream shopping**
X Liu - Marketing Science, 2023 - pubsonline.informs.org
... This paper presents a new solution to the dynamic **coupon targeting** problem based on batch **deep reinforcement learning**, and we conduct a large-scale field experiment to ...
☆ Save 99 Cite Cited by 67 Related articles All 5 versions Web of Science: 16 »

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf

informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–22
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Develops batch deep reinforcement learning (BDRL) to personalize dynamic coupon targeting.

Uses DR method to provide in-sample policy evaluation of the proposed BDRL policy.

Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping

Xiao Liu[a]
[a]Stern School of Business, New York University, New York, New York 10012
Contact: xl23@stern.nyu.edu, https://orcid.org/0000-0002-7093-8534 (XL)

Received: September 3, 2020
Revised: September 5, 2021; April 17, 2022
Accepted: June 23, 2022
Published Online in Articles in Advance: October 20, 2022
https://doi.org/10.1287/mksc.2022.1403

Abstract. We present an empirical framework for creating dynamic coupon targeting strategies for high-dimensional and high-frequency settings, and we test its performance using a large-scale field experiment. The framework captures consumers' intertemporal tradeoffs associated with dynamic pricing and does not rely on functional form assumptions about consumers' decision-making processes. The model is estimated using batch deep reinforcement learning (BDRL), which relies on Q-learning, a model-free solution that can mitigate model bias. It leverages deep neural networks to represent the high-dimensional state space and alleviate the curse of dimensionality. The empirical application is in a multibillion-dollar livestream shopping context. Our BDRL solution increases the platform's revenue by twice as much as static targeting policies and by 20% more than the model-based solution. The comparative advantage of BDRL comes from more effective and automatic targeting of consumers based on both heterogeneity and dynamics, using exceptionally rich, nuanced differences among consumers and across time. We find that price skimming, reducing discounts for attractive hosts, and increasing the coupon discount level at a faster rate for low spenders are effective strategies based on dynamics, consumer heterogeneity, and the two combined, respectively.

History: K. Sudhir served as the senior editor and John Hauser served as associate editor for this article.
Funding: Partial financial support was received from the NYU Center for Global Economy and Business.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mksc.2022.1403.

Keywords: dynamic pricing • coupon • deep reinforcement learning • reference price • livestream shopping • targeting

Total expected reward of policy π_e ← $V^{\pi_e} = E_n\left[\sum_{t=0}^{T-1}\delta^t\left(\underbrace{\xi_{0:t}[R_t - m_t(\mathbf{S}_t, A_t)]}_{\text{Importance Sampling: low bias}} + \underbrace{\xi_{0:t-1}\{\Sigma_{A\in\mathcal{A}} m_t(\mathbf{S}_t, A)\pi_e(A \mid \mathbf{S}_t)\}}_{\text{Direct Method: low variance}}\right)\right]$

where $\xi_{0:t} = \Pi_{\tau=0}^{t}\frac{\pi_e(A_\tau|\mathbf{S}_\tau)}{\pi_b(A_\tau|\mathbf{S}_\tau)}$ are the inverse propensity weights, and $\pi_b(A_\tau \mid \mathbf{S}_\tau)$ is the behavioral policy. In our case, the behavioral policy is known because the platform used a random allocation policy with predetermined action probabilities; $m_t(\mathbf{S}_t, A_t)$ is the direct method estimator, and for the functional form, we chose GBDT in which the Q value is the dependent variable and the state action variables $(\mathbf{S}, A)$ are the independent variables.

DR for Policy Learning

Targeting for **long-term outcomes**
J Yang, D Eckles, P Dhillon, S Aral - Management Science, 2024 - pubsonline.informs.org
... **outcomes** to optimize a **targeting** policy. We estimate the missing **long-term outcome** as the expectation of the **long-term outcome** conditional on surrogates of that **outcome** in a historical ...
☆ Save 99 Cite Cited by 80 Related articles All 11 versions Web of Science: 3 »

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–15
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Statistical surrogacy + policy learning to target for long-term outcomes.

Targeting for Long-Term Outcomes

Jeremy Yang,[a,*] **Dean Eckles,**[b,*] **Paramveer Dhillon,**[c] **Sinan Aral**[b]
[a]Harvard Business School, Boston, Massachusetts 02163; [b]Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [c]University of Michigan, Ann Arbor, Michigan 48109
*Corresponding authors
Contact: jeryang@hbs.edu, https://orcid.org/0000-0001-8639-5493 (JY); eckles@mit.edu, https://orcid.org/0000-0001-8439-442X (DE); dhillonp@umich.edu, https://orcid.org/0000-0002-0994-9488 (PD); sinan@mit.edu, https://orcid.org/0000-0002-2762-058X (SA)

Received: October 7, 2020
Revised: February 12, 2022
Accepted: February 27, 2022
Published Online in Articles in Advance: August 3, 2023
https://doi.org/10.1287/mnsc.2023.4881

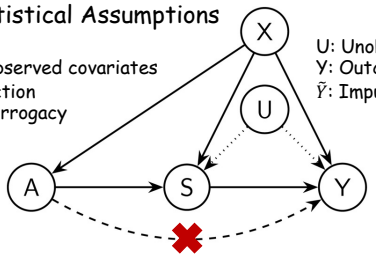
Abstract. Decision makers often want to target interventions so as to maximize an outcome that is observed only in the long term. This typically requires delaying decisions until the outcome is observed or relying on simple short-term proxies for the long-term outcome. Here, we build on the statistical surrogacy and policy learning literatures to impute the missing long-term outcomes and then approximate the optimal targeting policy on the imputed outcomes via a doubly robust approach. We first show that conditions for the validity of average treatment effect estimation with imputed outcomes are also sufficient for valid policy evaluation and optimization; furthermore, these conditions can be somewhat relaxed for policy optimization. We apply our approach in two large-scale proactive churn management experiments at *The Boston Globe* by targeting optimal discounts to its digital subscribers with the aim of maximizing long-term revenue. Using the first experiment, we evaluate this approach empirically by comparing the policy learned using imputed outcomes with a policy learned on the ground-truth, long-term outcomes. The performance of these two policies is statistically indistinguishable, and we rule out large losses from relying on surrogates. Our approach also outperforms a policy learned on short-term proxies for the long-term outcome. In a second field experiment, we implement the optimal targeting policy with additional randomized exploration, which allows us to update the optimal policy for future subscribers. Over three years, our approach had a net-positive revenue impact in the range of $4–$5 million compared with the status quo.

History: Accepted by Eric Anderson, marketing.
Funding: This work was supported by Boston Globe Media.
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2023.4881.

Keywords: long-term effect • statistical surrogate • policy learning • targeting • proactive churn management

Statistical Assumptions

X: Observed covariates
A: Action
S: Surrogacy



U: Unobserved covariates
Y: Outcomes
$\tilde{Y}$: Imputed Y based on S

$$\hat{\tilde{V}}_{\mathrm{DR}}(\pi_P) = \frac{1}{n}\sum_i\left(\hat{\tilde{\mu}}(\boldsymbol{X}_i, \pi_P) + \frac{\pi_P(A_i|\boldsymbol{X}_i)}{\pi_D(A_i|\boldsymbol{X}_i)}\cdot(\tilde{Y}_i - \hat{\tilde{\mu}}(\boldsymbol{X}_i, A_i))\right)$$

$$\hat{\tilde{\gamma}}_a(\boldsymbol{X}_i) = \hat{\tilde{\mu}}(\boldsymbol{X}_i, a) + \frac{\tilde{Y}_i - \hat{\tilde{\mu}}(\boldsymbol{X}_i, a)}{\pi_D(a|\boldsymbol{X}_i)}\cdot 1_{\{A_i=a\}}$$

$$\hat{\tilde{\pi}}^* = \underset{\pi\in\Pi}{\mathrm{argmax}}\ \frac{1}{n}\sum_i\left(\hat{\tilde{\gamma}}_1(\boldsymbol{X}_i) - \hat{\tilde{\gamma}}_0(\boldsymbol{X}_i)\right)\cdot(2\pi(\boldsymbol{X}_i) - 1)$$