DOTE 6635: Artificial Intelligence for Business Research

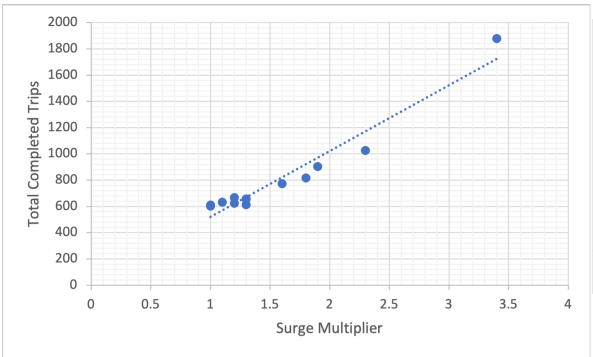
Causal Inference Fundamentals

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

		Condition: Mild	Condition: Severe	Condition: Total
Treatment	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,

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

- 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!

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

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

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

Agenda

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

17

17

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

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

18

18

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

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

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

19

19

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 $x \in \mathcal{X}$

20

20

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.

21

21

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

22

22

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:

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

23

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.

24

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:

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

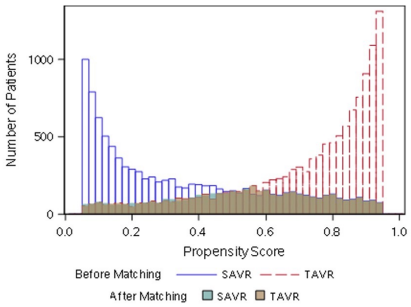
25

25

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.



26

26

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

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

27

27

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.*

28

28

Agenda

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

29

29

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 »

30

30

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: 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.

31

31