CI = UNCONFOUNDEDNESS

- Consider an RCT with $\{Y_i(0), Y_i(1)\} \perp W_i$

Potential outcomes with a population model $(Y_i(0), Y_i(1)) \sim \mathbb{P}$

$\sim \mathcal{B}(\pi)$ $\in \{0,1\}$

for $i = 1, \dots, n$ units, $Y_i = W_i Y_i(1) + (1-W_i) Y_i(0)$, $n_1 = \sum_{i=1}^{n} W_i$ and $n_0 = n - n_1$.

- In CI: RANDOMIZED CONTROL TRIALS (p. 4), we introduced

$$\boxed{\hat{\Delta} = \frac{1}{n_1} \sum_{i=1}^{n} W_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1-W_i) Y_i}$$

(DIFFERENCE ESTIMATOR)

as a natural estimator of the ATE $\Delta^\infty = \mathbb{E}(Y_i(1) - Y_i(0))$.

- Alternatively, noting that $\mathbb{E}\big(\underbrace{\sum_{i=1}^{n} W_i}_{n_1}\big) = n\pi$ and $\mathbb{E}\big(\underbrace{\sum_{i=1}^{n}(1-W_i)}_{n_0}\big) = n(1-\pi)$, we may replace n_0 (resp. n_1) in $\hat{\Delta}$ by $n(1-\pi)$ (resp. $n\pi$) and define:

$$\boxed{\tilde{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \frac{W_i Y_i}{\pi} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1-W_i) Y_i}{1-\pi}}$$

A natural estimator of Δ^∞ to consider as well.

= the Inverse Probability Weighting (IPW) estimator of Δ^∞ in a Bernoulli RCT.

As we will see in the next sections, IPW-like estimators play a central role when estimating causal effects in observational studies (under unconfoundedness) ↓

I - UNCONFOUNDEDNESS

[From Wager (2020)]

The simplest generalization of one RCT = multiple RCT

Suppose that we have $x = 1, \dots, K$ RCT i.e.

$$\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i \qquad (*)$$

For example, letting $x =$ 'Paris', 'Berlin', 'Melbourne', condition (*) means that three RCTs were conducted in these three different cities.

We write $e(x) = \mathbb{P}(W=1 \mid X = x)$

If the same Bern(π) randomization scheme was used in each of the three cities, then $e(x) = \pi$ for $x =$ 'Paris', 'Berlin' and 'Melbourne'.

We still want $\Delta^\infty = \mathbb{E}\{Y_i(1) - Y_i(0)\}$.

- $\forall x \in \mathcal{X} = \{1, \dots, K\}$, let

$\forall x$, consider the diff in means estimator $\hat{\Delta}(x)$

$$\hat{\Delta}(x) = \frac{1}{n_{x_1}} \sum_{\substack{W_i=1 \\ X_i = x}} Y_i - \frac{1}{n_{x_0}} \sum_{\substack{W_i=0 \\ X_i=x}} Y_i$$

$$\hat{\Delta}_{ADM} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\Delta}(x)$$

$n_x = n_{x_1} + n_{x_0}$

$n = \sum_{x \in \mathcal{X}} n_x$

Agg. Diff. in Means = (ADM)

Put $\pi_x = \mathbb{P}(X_i = x)$

$\hat{\pi}_x = n_x / n$

- Q: How accurate is $\hat{\Delta}_{ADM}$?

Put $\Delta^\infty = \sum_{x \in \mathcal{X}} \pi_x \Delta(x)$; $\Delta(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X = x\}$ (the Conditional ATE)(CATE)

Then

$$\hat{\Delta}_{ADM} - \Delta^\infty = \boxed{\sum_{x \in \mathcal{X}} \pi_x (\hat{\Delta}(x) - \Delta(x))} + \boxed{\sum_{x \in \mathcal{X}} (\hat{\pi}_x - \pi_x)\Delta(x)} + \boxed{\sum_{x \in \mathcal{X}} (\hat{\pi}_x - \pi_x)(\hat{\Delta}(x) - \Delta(x))} = O_p(1/n)$$

↘ Black term. Assuming $\text{var}(Y_i(w) \mid X_i = x) = \sigma^2(x)$ $(w = 0, 1)$,

then $n_x^{1/2}(\hat{\Delta}(x) - \Delta(x)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x)}{e(x)(1-e(x))}\right)$

$$\Rightarrow \sum_x \pi_x (\hat{\Delta}(x) - \Delta(x)) \approx \mathcal{N}\left(0, \frac{1}{n} \sum_x \pi_x^2 \frac{1}{\pi_x} \frac{\sigma^2(x)}{e(x)(1-e(x))}\right)$$

The asymptotic variance is

$$\sum_x \pi_x \frac{\sigma^2(x)}{e(x)(1-e(x))} = \mathbb{E}\left\{\frac{\sigma^2(X)}{e(X)(1-e(X))}\right\}$$

↘ Blue term $= \sum_x \frac{n_x}{n} \Delta(x) - \sum_x \Delta(x) \mathbb{P}(X_i = x)$

$= \frac{1}{n} \sum_x n_x \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x) - \mathbb{E}\{\Delta(X)\}$

$= \frac{1}{n} \sum_{i=1}^n \Delta(X_i) - \mathbb{E}\{\Delta(X)\}$

↳ CLT $\approx \mathcal{N}\left(0, \frac{\text{var}\,\Delta(X)}{n}\right)$

Putting pieces together:

$$n^{1/2}(\hat{\Delta}_{ADM} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V_{ADM})$$

where

$$V_{ADM} = \text{Var}\{\Delta(X)\} + \mathbb{E}\left\{\frac{\sigma^2(X)}{e(X)(1-e(X))}\right\}$$

↑ $X \in \{1, \dots, K\}$ has discrete support

What about X_i with continuous support?

→ $\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i$ is called UNCONFOUNDEDNESS introduced by Rosenbaum & Rubin (1984)

→ $e(x) = \mathbb{P}(W_i = 1 \mid X_i = x)$ is called the PROPENSITY SCORE (PS)

× Remark: Unconfoundedness $\Rightarrow \{Y_i(0), Y_i(1)\} \perp W_i \mid e(X_i)$ (see p. 64/65 in CI: ELEMENTS OF CAUSAL INFERENCE)

Algorithmic implication: cut into buckets along $e(X_i)$ and use $\hat{\Delta}_{ADM}$ (even though there is still a bit of confounding in each of the buckets).

II. INVERSE PROBABILITY WEIGHTING (IPW)

A conceptual implication of the previous remark is that the propensity score plays a central role. There are many ways to use PS for ATE estimation. One way is IPW.

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

= oracle (Horvitz-Thompson)

Compare with $\tilde{\Delta}$ on page 1 defined in the context of a single RCT: π is replaced by the PS $e(X_i) = \mathbb{P}(W_i = 1 \mid X_i = 1)$, a consequence of $\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i = x$

$\hat{\Delta}^*_{IPW}$ is an oracle estimator since the PS is assumed to be known. In practice, we consider

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

where $\hat{e}(\cdot)$ is an estimate of $e(\cdot)$.

× Result: $\mathbb{E}(\hat{\Delta}^*_{IPW}) = \Delta^\infty$

proof:

$$\mathbb{E}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

$$= \mathbb{E}\left(\frac{W_iY_i(1)}{e(X_i)} - \frac{(1-W_i)Y_i(0)}{1-e(X_i)}\right) \quad \text{(SUTVA)}$$

$$= \mathbb{E}\,\mathbb{E}\left(\text{---}''\text{---} \mid X_i\right)$$

unconfoundedness

$$= \mathbb{E}\left\{\frac{\cancel{\mathbb{E}(W_i \mid X_i)}\,\mathbb{E}(Y_i(1)\mid X_i)}{\cancel{e(X_i)}} - \dots\right\}$$

$$= \mathbb{E}\{Y_i(1) - Y_i(0)\}$$

$$= \Delta^\infty$$

∎

In addition, $n^{1/2}(\hat{\Delta}^*_{IPW} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V_{IPW})$

where $V_{IPW} = \mathrm{var}\left[\left(\frac{W_i}{e(X_i)} - \frac{1-W_i}{1-e(X_i)}\right)Y_i\right]$ since the oracle estimator $\hat{\Delta}^*_{IPW}$ is a sum of iid random variables. We derive next an alternative expression for V_{IPW}:

Put
- $\mathbb{E}(Y_i(w) \mid X_i) = \mu_{(w)}(X_i)$
- $\mathrm{var}(Y_i(w) \mid X_i) = \mathrm{var}(\varepsilon_i(w) \mid X_i) = \sigma^2(X_i)$
- $c(x) := e(x)\,\mu_{(0)}(x) + (1-e(x))\,\mu_{(1)}(x)$

just for convenience; it will simplify calculations of V_{IPW}.

It allows us to write

$$\begin{cases} Y_i(0) = \overbrace{\boxed{c(X_i) - (1-e(X_i))\,\Delta(X_i)}}^{\mu_{(0)}(X_i)} + \varepsilon_i(0) \\ Y_i(1) = \underbrace{\boxed{c(X_i) + e(X_i)\,\Delta(X_i)}}_{\mu_{(1)}(X_i)} + \varepsilon_i(1) \end{cases}$$

$\Delta(X_i) = \mu_{(1)}(X_i) - \mu_{(0)}(X_i)$
$(= \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i))$

$\Downarrow$

$$V_{IPW} = \mathrm{var}\left\{\left(\frac{W_i}{e(X_i)} - \frac{1-W_i}{1-e(X_i)}\right)c(X_i) + \Delta(X_i) + \left(\frac{W_i\,\varepsilon_i(1)}{e(X_i)} + \frac{(1-W_i)\,\varepsilon_i(0)}{1-e(X_i)}\right)\right\}$$

these three terms are uncorrelated

(7)

To see this, note that

$$\mathbb{E}\left\{\left(\frac{\omega}{e(X)} - \frac{1-\omega}{1-e(X)}\right) c(X) \mid X\right\} = 0 .$$

Thus

$$\begin{aligned}
\operatorname{Cov}&\left(\left(\frac{\omega}{e(X)} - \frac{1-\omega}{1-e(X)}\right) c(X), \Delta(X)\right) \\
&= \mathbb{E}\left\{\left(\frac{\omega}{e(X)} - \frac{1-\omega}{1-e(X)}\right) c(X)\, \Delta(X)\right\} \\
&= 0 \quad \text{(again, conditioning on } X \text{ first)}
\end{aligned}$$

The other two covariance terms are treated similarly (making use of $\mathbb{E}(\varepsilon(\omega) \mid X) = 0$).

$$\begin{aligned}
V_{IPW} &= \operatorname{var}\{\Delta(X_i)\} + \mathbb{E}\left\{\frac{\sigma^2(X_i)}{e(X_i)(1-e(X_i))}\right\} \\
&\quad + \boxed{\mathbb{E}\,\mathbb{E}\left\{\left(\frac{W_i}{e(X_i)} - \frac{1-W_i}{1-e(X_i)}\right)^2 c^2(X_i) \mid X_i\right\}}
\end{aligned}$$

$$\mathbb{E}\,\mathbb{E}\left\{\left(\frac{W_i}{e(X_i)} - \frac{1-W_i}{1-e(X_i)}\right)^2 c^2(X_i) \mid X_i\right\} = \mathbb{E}\left\{\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right\}$$

× Summary: $n^{1/2}(\hat{\Delta}^*_{IPW} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V_{IPW})$

where

$$V_{IPW} = \operatorname{var}\{\Delta(X_i)\} + \mathbb{E}\left\{\frac{\sigma^2(X_i)}{e(X_i)(1-e(X_i))}\right\} + \mathbb{E}\left\{\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right\}$$

To understand how good this is, let's check how V_{IPW} compares with V_{ADM} in the case where X_i has a discrete (finite) support.

(8)

We immediatly see that

$$V_{ADM} = V_{IPW} - \underbrace{\mathbb{E}\left\{\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right\}}_{\geq 0}$$

$\Rightarrow$ $\hat{\Delta}_{ADM}$ has a smaller asymptotic variance than the oracle IPW estimator $\hat{\Delta}^*_{IPW}$. But there is an even more suprising fact:

$$\begin{aligned}
\hat{\Delta}_{ADM} &= \sum_{x \in \mathcal{X}} \hat{\pi}_x \left(\frac{1}{n_{x_1}} \sum_{\substack{W_i=1\\X_i=x}} Y_i - \frac{1}{n_{x_0}} \sum_{\substack{W_i=0\\X_i=x}} Y_i\right) \\
&= \frac{1}{n} \sum_{W_i=1} \frac{W_i Y_i}{(n_{x_1}/n_x)} - \frac{1}{n} \sum_{W_i=0} \frac{(1-W_i) Y_i}{(n_{x_0}/n_x)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{W_i Y_i}{\hat{e}(X_i)} - \frac{1}{n} \sum_{i=1}^{n} \frac{(1-W_i) Y_i}{1-\hat{e}(X_i)}
\end{aligned}$$

$$\hat{e}(x) = \frac{n_{x_1}}{n_x}$$

The ADM estimator is an IPW estimator with estimated PS! And it has a smaller asymptotic variance than the oracle estimator.

(the estimated PS corrects for local variability, the same way that we divide by n_1 and not by $n\pi$ in the RCT setting).

$\Rightarrow$ The IPW estimator is not "optimal". In the next section, we introduce a variant of the IPW estimator with better asymptotic properties.

× Remarks: (a) We established consistency and a CLT for the oracle estimator $\hat{\Delta}^*_{IPW}$. Assuming that

(i) we have OVERLAP $\eta \le e(x) \le 1-\eta \quad \forall x \in \mathcal{X}$

(ii) $\exists M$ s.t. $|Y_i| \le M$

(iii) $\sup_{x\in\mathcal{X}} |\hat{e}(x) - e(x)| = O_p(a_n) \to 0$,

then one can show that $|\hat{\Delta}_{IPW} - \hat{\Delta}^*_{IPW}| = O_p\left(\frac{a_n M}{\eta}\right)$.

In other words, under assumptions (i)-(ii)-(iii), $\hat{\Delta}^*_{IPW}$ is consistent $\Rightarrow$ $\hat{\Delta}_{IPW}$ is consistent.

To prove this, we use that $\hat{e}(x)$ becomes uniformly bounded away from 0 and 1 as $n\to\infty$ under (iii), which in turns implies that $1/\hat{e}(x)$ and $1/1-\hat{e}(x)$ decay at the same $O_p(a_n)$ rate as $\hat{e}(x)$.

(b) Rewriting $\underbrace{q(w,x)}_{\text{weight}} = 1/\underbrace{\mathbb{P}(W=w|X=x)}_{\text{P.S.}}$,

we see that $\mathbb{E}(W q(W,X)) = \mathbb{E}((1-W)q(W,X)) = 1$.

In addition,

$$\mathbb{E}Y_i(1) \underset{p.5}{=} \mathbb{E}\left(\frac{W_i Y_i}{e(X_i)}\right) = \mathbb{E}\left(W_i Y_i\, q(W_i, X_i)\right) = \frac{\mathbb{E}(W_i Y_i\, q(W_i,X_i))}{\underbrace{\mathbb{E}(W_i\, q(W_i,X_i))}_{=1}}$$

and similarly for $\mathbb{E}Y_i(0)$. This suggests defining a version

of the IPW estimator using normalized weights:

$$\tilde{\Delta}^*_{IPW} = \sum_{i|W_i=1} \left\{\frac{q(W_i,X_i)}{\sum_{j|W_j=1} q(W_j,X_j)}\right\} Y_i - \sum_{i|W_i=0} \left\{\frac{q(W_i,X_i)}{\sum_{j|W_j=0} q(W_j,X_j)}\right\} Y_i$$

$$= \sum_{i|W_i=1} q'(W_i,X_i)Y_i - \sum_{i|W_i=0} q'(W_i,X_i)Y_i$$

(Robins '98)

Normalized weights are often preferred over original weights, especially when $e(x)$ is close to 0 or 1, which increases the variance of the IPW estimator.

III_ BALANCING WEIGHTS

IPW is a special case of a general class of balancing weights. We introduced p.3 the CATE

$$\Delta(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X = x\}$$

Usually, the CATE is not computed for a single x, but averaged over a target distribution of the covariates. Let

↘ $f(x)$ = marginal density of X over the whole pop.

↘ $g(x)$ = density of X of a target population, possibly different from X.

Ex: Medical treatment applied to the whole population ($X\sim f$) vs Medical treatment applied to those individuals who need it the most ($X \sim g$)

↘ $h(x) = \dfrac{g(x)}{f(x)}$ reweights observations to represent the target population.

We introduce next a new class of estimators: the ATE over the target population g (Li, Morgan, Zaslavsky 2018)

$$\Delta_h^\infty := \mathbb{E}_g\left[Y_i(1) - Y_i(0)\right] = \frac{\mathbb{E}\{h(X)\Delta(X)\}}{\mathbb{E}\{h(X)\}} \quad X \sim f$$

$$= \frac{\int \Delta(x) f(x) h(x)\, dx}{\int f(x) h(x)\, dx}$$

Let $f_w(x) = \mathbb{P}(X = x \mid W = w)$ (working w.l.o.g. with discrete distributions here)

$=$ density of X in the $W = w$ group.

Then

$$f_1(x) = \frac{\mathbb{P}(X = x, W = 1)}{\mathbb{P}(W = 1)} = \frac{\mathbb{P}(W = 1 \mid X = x)\,\mathbb{P}(X = x)}{\mathbb{P}(W = 1)} \propto f(x)\, e(x)$$

and similarly $f_0(x) \propto f(x)(1 - e(x))$

For a specific $h(x)$, we can estimate Δ_h^∞ using weights $w_0(x)$, $w_1(x)$ defined such that

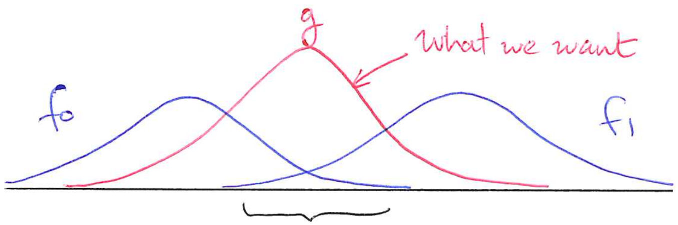
$$\underset{\uparrow}{w_0(x)}\, f_0(x) = \underset{\uparrow}{w_1(x)}\, f_1(x) = \underbrace{h(x) f(x)}_{\text{target } g(x)}$$

reweight each $W = w$ group

$$\Rightarrow \begin{cases} w_0(x) \propto \dfrac{f(x)h(x)}{f(x)(1 - e(x))} = \dfrac{h(x)}{1 - e(x)} \\[2ex] w_1(x) \propto \dfrac{f(x)h(x)}{f(x)e(x)} = \dfrac{h(x)}{e(x)} \end{cases}$$

- Ex: $h(x) = 1$; the target population is $f(x)$; weights are $(w_0(x), w_1(x)) = \left(\dfrac{1}{1 - e(x)}, \dfrac{1}{e(x)}\right)$ and $\Delta_h^\infty = \Delta^\infty = \mathbb{E}\{Y_i(1) - Y_i(0)\}$ ↑ IPW
- Ex: $h(x) = e(x)$; the target population is $f(x)e(x) \propto f_1(x)$ i.e. the treated subpopulation. The weights are $(w_0(x), w_1(x)) = \left(\dfrac{e(x)}{1 - e(x)}, 1\right)$ and the estimand is $\Delta_h^\infty = \text{ATT} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid W = 1\}$
- Ex: $h(x) = 1 - e(x)$, the target population is the control subpopulation. The weights are $(w_0(x), w_1(x)) = \left(1, \dfrac{1 - e(x)}{e(x)}\right)$ and $\Delta_h^\infty = \text{ATNT} = \mathbb{E}\{Y_i(1) - Y_i(0) \mid W = 0\}$
- Ex: $h(x) = $ indicator function $= \mathbb{1}(\alpha < e(x) < 1 - \alpha)$, $\alpha \in (0, 1/2)$ leads to an ATE for a subpopulation with overlap of the covariates between the two groups.
- Ex: $h(x) = e(x)(1 - e(x))$ yields $(w_0, w_1) = (e(x), 1 - e(x))$. The estimand $\Delta_h^\infty = \mathbb{E}\{e(X)(1 - e(X))\Delta(X)\} / \mathbb{E}\{e(X)(1 - e(X))\}$ is called the ATE for the overlap population.

Indeed, $h(x)$ is maximal when $e(x) = 1/2$; ie. for those individuals which have an equal chance to be allocated to a treatment or control group.



grey area: in medical applications, unsure if this patient should be given this new treatment or not $\Rightarrow$ we want to focus on these patients the most, and this is exactly what $h(x)$ is doing.

IV_ AUGMENTED IPW

We refer to Chapter 3 in Wager (2020) for a formal account on AIPW.

BALANCING WEIGHTS

(From a keynote talk of Betsy Ogburn at EUROCIM '24)

[REF] Augmented balancing weights as linear regression
D. Bruns-Smith, O. Dukes, A. Feller & B. Ogburn.

- Consider iid tuples $(X_i, T_i, Y_i(0), Y_i(1))$, with $X_i \in \mathbb{R}^d$ (cov), $T_i \in \{0,1\}$ (trt alloc), $Y_i(0) \in \mathbb{R}$, $Y_i(1) \in \mathbb{R}$

↘ We are interested in estimating $\Delta = \mathbb{E}[Y(1) - Y(0)]$
& focus on the counterfactual mean $\mu_1 := \mathbb{E}[Y(1)]$.

- Usual assumptions → Ignorability $T \perp \{Y(0), Y(1)\} \mid X$
→ Overlap $e(x) = \mathbb{P}(T=1 \mid X=x) > 0$

- Two common strategies:

OUTCOME MODELLING

$$\mu(x,t) = \mathbb{E}(Y \mid X=x, T=t)$$

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

→ IPW

$$\mu_1 = \mathbb{E}\left(\frac{T}{e(X)} Y\right)$$

under ignorability

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

Doubly Robust estimators combine both approaches.

Ex: AIPW

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

Issues with inverse PS weights: small errors in estimating the PS are magnified when taking its inverse.
& obs with small proba of treatment dominate (expected, but again small errors are demultiplied).

Alternative approach = estimate the inverse PS directly.

Put $w(X) = \frac{1}{e(X)}$. We have =

$$\mu_1 = \mathbb{E}[Y(1)] = \mathbb{E}\,\mathbb{E}[Y(\cancel{1}) \mid X, T=1] = \mathbb{E}\,\mu(X,1)$$

$$= \mathbb{E}[T w(X) Y]$$

$$= \mathbb{E}[T w(X) \mu(X,1)]$$

$$\Downarrow$$

$$\boxed{\mathbb{E}\,\mu(X,1) = \mathbb{E}[T w(X) \mu(X,1)]}$$

In fact, we can show that $w(x) = \frac{1}{e(x)}$ is the only functional satisfying this equality for all measurable functions $m(\cdot, 1)$ [Riesz Representer]

We can use this property to characterize $w(x)$ as the unique solution to the following optimization problem

$$\min_{w} \sup_{f} \left\{ \mathbb{E}[T w(X) f(X)] - \mathbb{E}[f(X)] \right\}$$

measurable

In practice, restrict to a class $\mathcal{F}$ of functionals & derive the optimal weights balancing all functions in that class. Let:

$$\text{Imbalance}_{\mathcal{F}}(w) := \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[T w(X) f(X)] - \mathbb{E}[f(X)] \right\}$$

Introduce a regularization hyperparameter $\delta > 0$ to ensure a unique min & for bias-variance tradeoff.

$$w^*(.) = \underset{w}{\operatorname{argmin}} \left\{ \text{Imbalance}_{\mathcal{F}}(w) + \delta \|w\|^2 \right\}$$

δ typically selected via cross-validation.

x Guarantees = Hirshberg & Wager (2021) show that if $m(\cdot, 1) \in \mathcal{F}$, then imbalance is small & that $\mathbb{E}[T w^*(X) Y]$ can be used as an approx. unbiased estimate of μ_1.
[aka imbalance$_{\mathcal{F}}$ used to bound the estimator's bias]

Ogburn et al restrict analysis to the space of linear functionals $\mathcal{F} = \{ f(x) = \theta^t x \;,\; \theta, x \in \mathbb{R}^d \}$, $\|\theta\| \leq 1$

any norm on $\mathbb{R}^d$

Authors are more general & consider rich sets of covariates [& potentially ∞-dim space $\mathcal{F}$ using RKHS]

$$\text{Imbalance}_{\mathcal{F}}(w) = \sup_{\|\theta\| \leq 1} \theta^t [\mathbb{E}\, T w(X) X - \mathbb{E} X]$$

$$= \| \mathbb{E}\, T w(X) X - \mathbb{E} X \|_*$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Ex: ℓ_2 is the dual norm of ℓ_2
ℓ_∞ is the dual norm of ℓ_1.

↳ Consider ℓ_2 balancing.

The dual representation shows that we are minimising the ℓ_2 norm of the difference $\mathbb{E}\, T w(X) X - \mathbb{E} X$.
The sample balance property is

$$\sum_{i=1}^{n} w(X_i) T_i X_{ij} \approx \sum_{i=1}^{n} X_{ij} \qquad j = 1, -, d$$

$X = \begin{pmatrix} X_1 \\ | \\ X_d \end{pmatrix}$

Want: balancing for each component

covariate distribution of the tratment group

Covariate distribution of the target group (whole sample)

Let $\hat{w}^{\ell_2}_{\delta}(\cdot)$ denote the solution to the (sample) optimization problem. The DR estimator of μ_1 is (5)

(*)

$$\frac{1}{n}\sum_{i=1}^{n}\left\{ \hat{\mu}^{RR}_{\lambda}(X_i,1) + \hat{w}^{\ell_2}_{\delta}(X_i)T_i\left(Y_i - \hat{\mu}^{RR}_{\lambda}(X_i,1)\right)\right\}$$

Outcome model fit using RR with penalty λ

$\hat{\mu}^{RR}_{\lambda}(x,1) = x^t \hat{\beta}^{RR}_{\lambda}$

↳ regularization when d is large

↳ bias/variance tradeoff

issue = we bring confounding back ⇒ RR does not perform well in high dim settings for estimating a causal effect in the presence of confounding

Augmented Term = Bias Correction

Required when the outcome model is regularized.

The main result of Ogburn et al is to show that in this setting [RR(λ) for outcome model + ℓ_2 norm with hyp δ for w],

this expression collapses to a form that can be represented as a single output regression

$$(*) = \frac{1}{n}\sum_{i=1}^{n} X_i^t \hat{\beta}^{AUG}_{\lambda,\delta}$$

In fact, $\hat{\beta}^{AUG}_{\lambda,\delta}$ is the solution of a (generalized) ridge regression. (see later)

To further characterize the solution of the generalized RR, we consider an important special case: we compute $\hat{w}^{\ell_2}_{0}(\cdot)$ with exact balancing. (6)

↳ The most common optimization problem in that case is

$$\min_{w\in\mathbb{R}^n} \|w\|_2^2$$

more unknowns than eq when $d<n$

$$\text{such that } \sum_{i=1}^{n} w_i T_i X_{ij} = \sum_{i=1}^{n} X_{ij} \qquad j=1,\dots,d$$

The Lagrangian of this convex opt. problem is

$$\mathcal{L}(w,v) = \sum_{i=1}^{n} w_i^2 + \sum_{j=1}^{d} v_j\left(\sum_{i=1}^{n} X_{ij} - \sum_{i=1}^{n} w_i T_i X_{ij}\right)$$

KKT conditions: optimal (w^*, v^*) parameters must satisfy:

- primary constraints $\sum_{i=1}^{n} X_{ij} = \sum_{i=1}^{n} w_i^* \boxed{T_i X_{ij}}$ $/\!/ \ \tilde{X}_{ij}$
- $\nabla_w \mathcal{L}(w,v) = 0 \Rightarrow 2w_i^* = \sum_{j=1}^{d} v_j \boxed{T_i X_{ij}}$ $\quad i=1,\dots,n$

Matrix notation

$$\underset{(n\times d)}{X} = \begin{pmatrix} — X_1^t — \\ \vdots \\ — X_n^t — \end{pmatrix} \quad \underset{(n\times d)}{\tilde{X}} = \begin{pmatrix} — \tilde{X}_1^t — \\ \vdots \\ — \tilde{X}_n^t — \end{pmatrix} \quad v^* = \begin{pmatrix} v_1^* \\ \vdots \\ v_d^* \end{pmatrix} \quad w^* = \begin{pmatrix} w_1^* \\ \vdots \\ w_n^* \end{pmatrix}$$

$$\mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n \qquad \text{KKT} \begin{cases} 2w^* = \tilde{X} v^* \\ X^t \mathbb{1} = \tilde{X}^t w^* \end{cases}$$

$$\Rightarrow v^* = 2(\tilde{X}^t\tilde{X})^{-1} X^t \mathbb{1}$$

$$\underset{(n\times 1)}{w^*} = \tilde{X}(\tilde{X}^t\tilde{X})^{-1} X^t \mathbb{1} \quad \leftarrow \text{linear in } X$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}^{RR}_{0}(X_i,1) = \frac{1}{n}(w^*)^t Y = \frac{1}{n}\mathbb{1}^t X(\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t Y$$

7

× <u>Remarks</u>: (i) The weighting estimator of μ_1 that exactly balances features is numerically equal to the OLS linear reg. estimator.

Indeed, assuming $\mu(x,1) = x^t\beta$,

$$\hat\beta^{OLS} \leftarrow OLS(Y_i \sim X_i \mid T_i = 1)$$

Fit one model for the trt group & one for the control group, separately.

$$\hat\beta^{OLS} = (\tilde X^t \tilde X)^{-1} \tilde X^t Y$$

restricting the sample to $T_i = 1$ or truncating entries in X and using $\tilde X$ yield the same numerical result.

$\Rightarrow$ Apply $\hat\beta^{OLS}$ to the whole (target) pop:

$X\hat\beta^{OLS} = X(\tilde X^t \tilde X)^{-1}\tilde X^t Y$, and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \hat\beta^{OLS} = \frac{1}{n} 1^t X(\tilde X^t \tilde X)^{-1} \tilde X^t Y \quad \blacksquare$$

(ii) This justifies the notation $\hat\beta^{OLS} = \hat\beta_0^{RR}$.

(iii) This result is long known, see Robins et al (2007)

$\Rightarrow$ When both the reg function and the weights are linear, the OLS estimator is DR.

Back to our problem, Ogburn et al show that $\hat\beta^{AUG}_{\lambda,\delta}$ [under a ridge penalty (λ) for the outcome model and a ridge penalty (δ) for the weights] can be expressed as a shrank version of $\hat\beta^{OLS}$. Specifically,

8

$$\left.\begin{matrix} \hat\beta^{RR}_{\lambda,j} = \left(\dfrac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)\hat\beta^{OLS}_j \\ + \\ \hat w^{\ell_2}_\delta \end{matrix}\right\} \Rightarrow \hat\beta^{AUG}_j = \left(\frac{\sigma_j^2}{\sigma_j^2 + \gamma_j}\right)\hat\beta^{OLS}_j$$

with

$$\gamma_j = \frac{\delta\lambda}{\sigma_j^2 + \lambda + \delta}$$

Under the assumption that $\tilde X^t \tilde X$ is diagonal $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$

Close form expression

Ridge(λ) + Ridge(δ) = Ridge

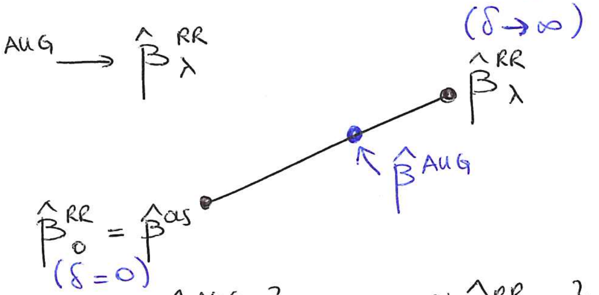
↘ $\lambda = 0$ [OLS] $\Rightarrow \hat\beta^{AUG} = \hat\beta^{OLS}$ Robins et al (2007)

↘ $\delta = 0 \Rightarrow \hat\beta^{AUG} = \hat\beta^{OLS}$

⚠ In practice, the Ridge + Ridge procedure with cross validation spits out the OLS estimator, since $\delta = 0$ is selected.

↳ Most researches are unaware that they are simply doing OLS

↘ $\delta \to \infty \Rightarrow \hat\beta^{AUG} \longrightarrow \hat\beta^{RR}_\lambda$



Also, can show that $\|Y - \tilde X\hat\beta^{AUG}\|_2^2 \le \|Y - \tilde X\hat\beta^{RR}_\lambda\|_2^2$

$\Rightarrow \hat\beta^{AUG}$ has smaller in-sample training error than $\hat\beta^{RR}_\lambda$.

9

$\Rightarrow$ The bias correction term in the DR procedure brings the overall solution back closer to the OLS "undersmoothing" / "underfitting".

↳ the present framework automatically selects the amount of undersmoothing required for the estimator to achieve $\sqrt{n}$ convergence (not true for the Ridge estimator due to regualisation & optimization of the MSE).

In practice, the authors suggest to cross validate $\hat{\beta}^{RR}_{\lambda}$ to select λ, and then take $\delta = \lambda$ and $\gamma = \lambda^2$ (data driven & seems to work well in practice).