STA 640 — Causal Inference

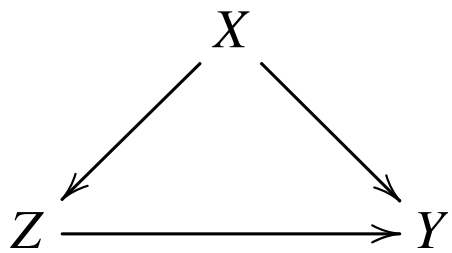
Chapter 3.4: Propensity Score Weighting

Fan Li

Department of Statistical Science
Duke University

Confounding

- Confounding X (or common cause; see DAG) is the main complication/hurdle between association and causation



- Propensity score methods: need a model for $Z \sim X$
  - Stratification
  - Matching
  - Regression (need additional model for $Y \sim e(X)$)
  - Weighting (this lecture)

Inverse Probability Weighting (IPW)

- Inverse probability weighting (IPW), also known as inverse probability of treatment weighting (IPTW)
- Easy to show

$$\mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right] = \tau^{\text{ATE}}.$$

Observe

$$\mathbb{E}\left[\frac{ZY}{e(X)}\right] = \mathbb{E}\left\{\frac{1}{e(X)}\,\mathbb{E}\left[ZY(1) \mid X\right]\right\} = \mathbb{E}\left\{\frac{\mathbb{E}(Z|X)}{e(X)}\,\mathbb{E}[Y(1)|X]\right\}$$

$$= \mathbb{E}\left\{\mathbb{E}[Y(1)|X]\right\} = \mathbb{E}[Y(1)]$$

- Horvitz-Thompson (HT) estimator in survey sampling literature

Inverse Probability Weighting (IPW)

- Recall

$$\mathbb{E}\left[\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right] = \tau^{\text{ATE}}.$$

- Define the inverse probability weights (IPW) :

$$\begin{cases} w_1(X_i) = \frac{1}{e(X_i)}, & \text{for } Z_i = 1 \\ w_0(X_i) = \frac{1}{1-e(X_i)}, & \text{for } Z_i = 0. \end{cases}$$

- An unbiased nonparametric (moment-based) estimator of ATE: difference in the mean of the weighted outcomes between groups

$$\begin{aligned} \hat{\tau}_{ipw,1} &= \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{Y_i Z_i}{e(X_i)} - \sum_{i=1}^{N}\frac{Y_i(1-Z_i)}{1-e(X_i)}\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{Y_i Z_i w_1(X_i) - Y_i(1-Z_i)w_0(X_i)\right\}. \end{aligned}$$

Normalize the weights

- When use any weighting method (e.g. IPW), good practice is to normalize weights – sum of the total of weights within one group should be 1
- Divide each unit's weight by the sum of all weights in that group $w_i / \sum_{i:Z=z} w_i$ for $z = 0, 1$, i.e. the Hajek estimator:

$$\hat{\tau}_{ipw,2} = \frac{\sum_{i=1}^{n} Y_i Z_i w_1(X_i)}{\sum_{i=1}^{n} Z_i w_1(X_i)} - \frac{\sum_{i=1}^{n} Y_i (1 - Z_i) w_0(X_i)}{\sum_{i=1}^{n} (1 - Z_i) w_0(X_i)}.$$

- Reduce variance, $\mathbb{V}(\hat{\tau}_{ipw,2}) \leq \mathbb{V}(\hat{\tau}_{ipw,1})$, and lead to more stable estimate (Hirano, Imbens, Ridder, 2003)
- Normalized weights are also called "stabilized weights" (Hernan, Robins, Brumback, 2000)

Inverse Probability Weighting (IPW)

- IPW creates a weighted population (i.e. target population) – a population that the study sample is representative of
- in this weighted population, the covariates distributions between two groups are balanced: can show

$$\mathbb{E}[XZ/e(X)] = \mathbb{E}[X(1-Z)/(1-e(X))].$$

- The intuition behind: the less likely a subject is sampled, then the larger population it should represent
- Up-weigh “weirdos": units whose treatment assignment is opposite to what the propensity score predicts

Inverse Probability Weighting (IPW): Variance

- What if the true PS $e(X)$ is known?
- Hirano, Imbens, Ridder (2003) shows that using the estimated PS instead of the true PS asymptotically gives more efficient estimate of treatment effect (projection)
- Closed-form sandwich estimator (M-estimator) of variance that takes into account of the uncertainty in estimating the propensity score (Lunceford and Davidian, 2004)
- Bootstrap: Resample units and refit PS and estimate the causal effects every time – computationally intensive for large sample

Inverse Probability Weighting (IPW)

Advantages (general to weighting)

- Simple, with theoretical foundation
- Smooth - explicit target population
- Global balance
- Extends to more complex settings

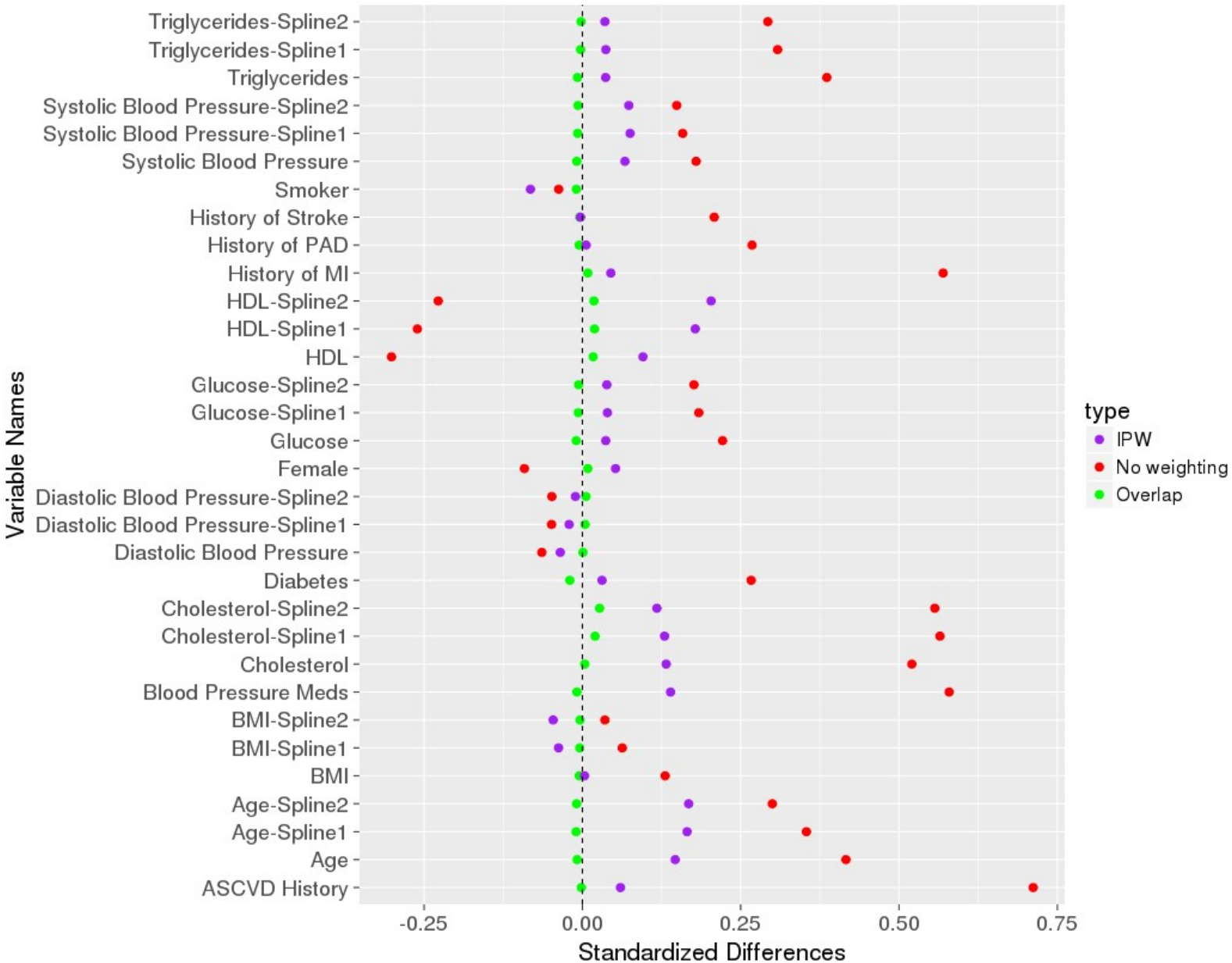
Disadvantages (specific to IPW)

- More sensitive to misspecification of propensity scores
- Propensity scores near 0 or 1 yield extreme weights
- Many ad hoc decisions to deal with this
- ATE might not always be the sensible estimand

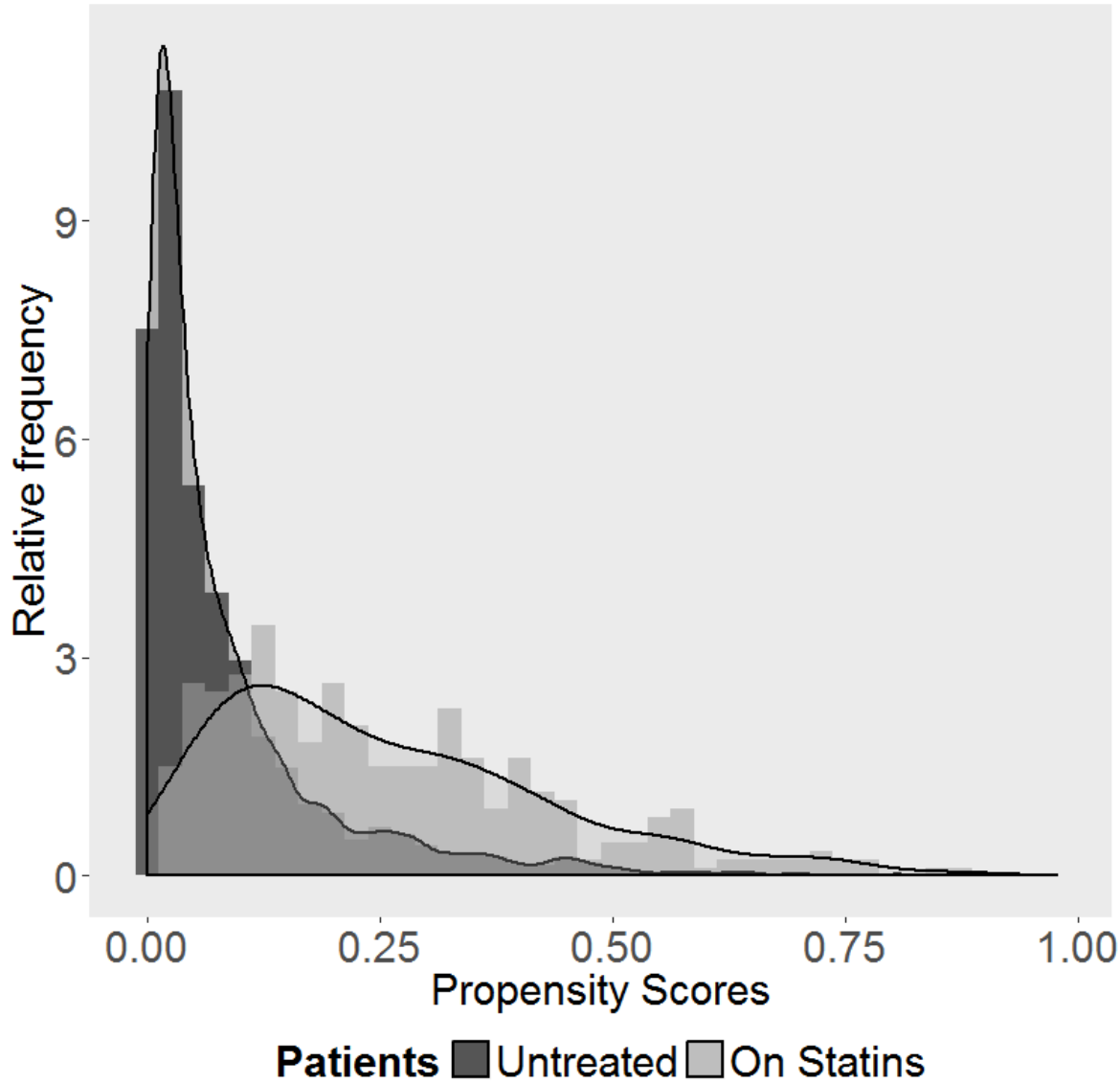
Example: Framingham Heart Study re-visited

- **Goal:** evaluate the effect of statins on health outcomes
- **Patients:** cross-sectional population from the offspring cohort with a visit 6 (1995-1998)
- **Treatment:** statin use at visit 6 vs. no statin use
- **Outcomes:** CV death, myocardial infarction (MI), stroke
- **Confounders:** sex, age, body mass index, diabetes, history of MI, history of PAD, history of stroke...
- Significant imbalance between treatment and control groups in covariates motivates IPW (or some form of propensity score adjustment)

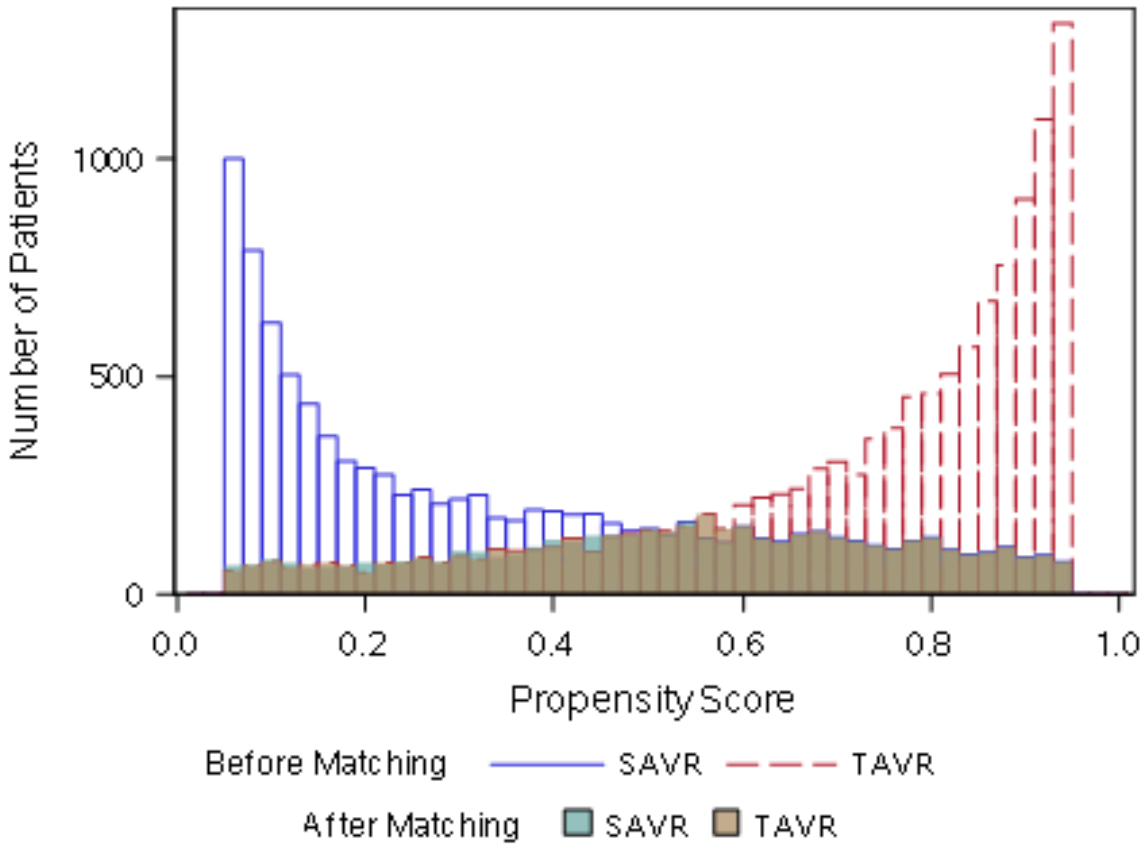
Balance in Framingham Study (Statins vs. Control)



Challenge to IPW: Poor overlap



Example: Poor Overlap (Brennen et al. 2016)



IPW Operational Challenges

- Propensity values near 0 and 1 yield extreme weights (after taking the inverse)
- Adverse finite-sample consequences – Basu's elephant: severe bias and variance
- Normalization of weights helps, but not a lot
- Core problem: lack of overlap in the tail of the propensity distribution – causal comparisons of these units are highly uncertain

Other adjustment methods

How do poor overlap and extreme propensity scores impact other adjustment methods?

- Regression adjustment: Increases model sensitivity
- Matching: Many patients get excluded because they don't have a match
- Stratification: Adds bias due to residual imbalance within strata

Better to fix IPW than to abandon weighting altogether.

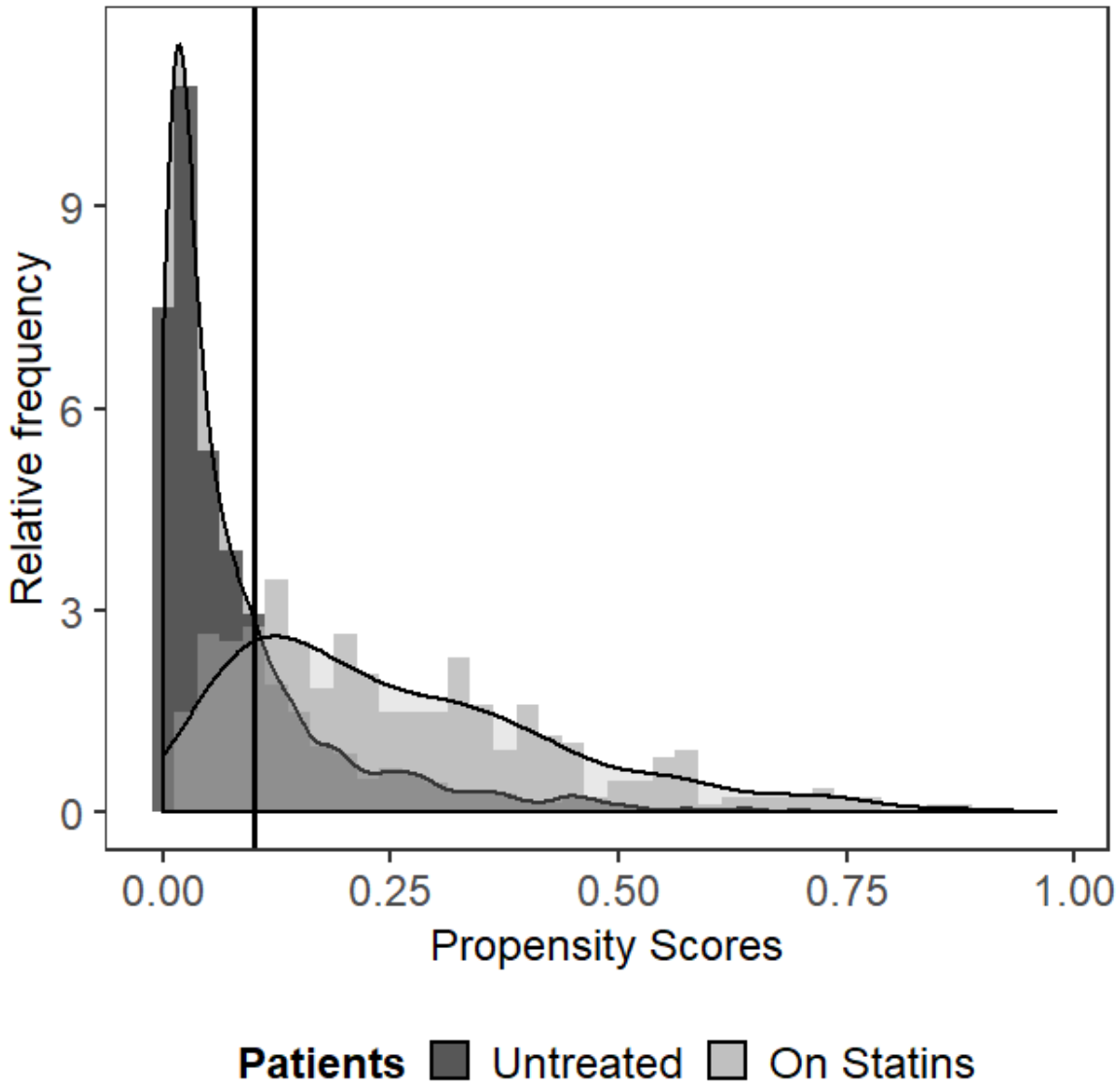
Propensity Score Trimming and Truncation

- Symmetric trimming (*Crump et al., 2009*)
  - exclude patients whose estimated PS is outside $[\alpha, 1-\alpha]$
  - rule of thumb $\alpha = 0.1$
- Asymmetric trimming (*Sturmer et al., 2010*)
  - exclude patients with PS outside of the common PS range formed by the treated and control patients
  - among all units, further exclude those whose PS is below the q quantile of the treated units
  - among all units, exclude those whose PS is above the $(1-q)$ quantile of the control units
- Propensity score truncation
  - Set the patients whose estimated PS is below α to α, and whose estimated PS is above $1-\alpha$ to $1-\alpha$
  - You can't be more ad hoc than this!

Trimming thresholds applied to Framingham

Method	Left Excluded	Right Excluded	Total
Symmetric $\alpha = 0.025$	1047	1	1048 (31%)
Symmetric $\alpha = 0.05$	1634	1	1635 (48%)
Symmetric $\alpha = 0.10$	2269	1	2270 (68%)
Asymmetric $q = 0.025$	1179	134	1313 (39%)
Asymmetric $q = 0.05$	1554	257	1811 (54%)
Asymmetric $q = 0.10$	1811	468	2279 (68%)

Symmetric Trimming $\alpha = 0.10$



Propensity Score Trimming - Cont'd

Reduce the impact of extreme PS and improve finite-sample property of IPW

Choice of threshold α, q may be arbitrary

- conceptual challenge: ambiguous target population/interpretation
- operational challenge: causal estimates sensitive to trimming threshold
- operational challenge: bias-variance tradeoff
- operational challenge: refitting PS after trimming (a hidden message)

Extension Beyond IPW: Move the Goalpost

- A different approach: move the goalpost away from ATE (IPW). Other, user-specified target populations.
- Key question: ATE over what target population? (Thomas et al. 2020a)
- Often units who are “one the fence" (in the middle of PS distribution) are of interest. Examples:
  1. TAVR/SAVR
  2. Under clinical equipoise, the target population are units with the most “overlap” observed characteristics.

Extension Beyond IPW: Move the Goalpost

- Motivating questions:
  1. Is there a unified framework for weighting for different target populations?
  2. Is there an intrinsic approach to define such "marginal population" while avoiding the pitfall of IPW?

Balancing Weights: A General Framework

Li, Morgan, Zaslavsky, 2018

- IPW is a special case of a general class of *balancing weights*
- Recall the conditional average treatment effect CATE is defined as

$$\tau(x) \equiv \mathbb{E}(Y(1)|X = x) - \mathbb{E}(Y(0)|X = x).$$

- 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

Balancing Weights: A General Framework

Li, Morgan, Zaslavsky, 2018

- A new class of estimands: the ATE over the target population g

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

- τ_h is a general class of weighted ATE (WATE) estimands (Hirano et al., 2003)
- When $h(x) = 1$, $f(x) = g(x)$, the target population is the observed population
- Varying $h(x)$ defines varying target population. In practice, we pre-specify the tilting function $h(x)$

Balancing Weights: A General Framework

Li, Morgan, Zaslavsky, 2018

- Let $f_z(x) = \Pr(X = x|Z = z)$ (covariate distribution within each group z), easy to show

$$f_1(x) \propto f(x)e(x), \quad f_0(x) \propto f(x)(1 - e(x))$$

- For a given $h(x)$, to estimate τ_h, we can weight $f_z(x)$ to the target population $g(x) = f(x)h(x)$ using analytic weights

$$\begin{cases} w_1(x) \propto \frac{f(x)h(x)}{f_1(x)} = \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, \\ w_0(x) \propto \frac{f(x)h(x)}{f_0(x)} = \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)}. \end{cases} \tag{2}$$

- The class of weights (w_0, w_1) is called balancing weights: balance the distributions of the weighted covariates between comparison groups $f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)h(x)$

Examples: target population (h) and balancing weights

- Choice of $h(x)$ determines the target population, estimand, weights
- Statistical, scientific and policy considerations all come into play in selecting $h(x)$

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	$(1-e(x), e(x))$
trimming	$\mathbf{1}(\alpha \le e(x) \le 1-\alpha)$		$\left(\frac{\mathbf{1}(\alpha \le e(x) \le 1-\alpha)}{e(x)}, \frac{\mathbf{1}(\alpha \le e(x) \le 1-\alpha)}{1-e(x)}\right)$
matching	$\min\{e(x), 1-e(x)\}$		$\left(\frac{\min\{e(x),1-e(x)\}}{e(x)}, \frac{\min\{e(x),1-e(x)\}}{1-e(x)}\right)$

Large-sample properties

- Sample estimator of WATE

$$\hat{\tau}_h = \frac{\sum_i w_1(x_i) Z_i Y_i}{\sum_i w_1(x_i) Z_i} - \frac{\sum_i w_0(x_i)(1 - Z_i) Y_i}{\sum_i w_0(x_i)(1 - Z_i)} \tag{3}$$

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

Large-sample properties

Theorem 2. *As $N \to \infty$, the expectation (over possible samples of covariate values) of the conditional variance of the estimator $\hat{\tau}_h$ given the sample $X = \{x_1, \ldots, x_N\}$ converges:*

$$N \cdot \mathbb{E}_x \, \mathbb{V}[\hat{\tau}_h \mid X] \quad \to \quad \int f(x) h(x)^2 \left[\frac{v_1(x)}{e(x)} + \frac{v_0(x)}{1 - e(x)} \right] \mu(dx) \Big/ C_h^2,$$

where $v_z(x) = \mathbb{V}[Y(z) \mid X]$ and $C_h = \int h(x) f(x) d\mu(x)$ is a normalizing constant.

Corollary 1. *The function $h(x) \propto e(x)(1 - e(x))$ gives the smallest asymptotic variance for the weighted estimator $\hat{\tau}_h$ among all h's under homoscedasticity, and as $N \to \infty$,*

$$N \cdot \min_h \{\mathbb{V}[\hat{\tau}_h]\} \to v \Big/ \int f(x) e(x)(1 - e(x)) \mu(dx).$$

Overlap Weights

- Based on Corollary 1, LMZ propose a new type of weights, the overlap weights, by letting $h(x) = e(x)(1 - e(x))$:

$$\begin{cases} w_1(x) \propto 1 - e(x), & \text{for } Z = 1, \\ w_0(x) \propto e(x), & \text{for } Z = 0. \end{cases}$$

- Each unit is weighted by its probability of being assigned to the opposite group

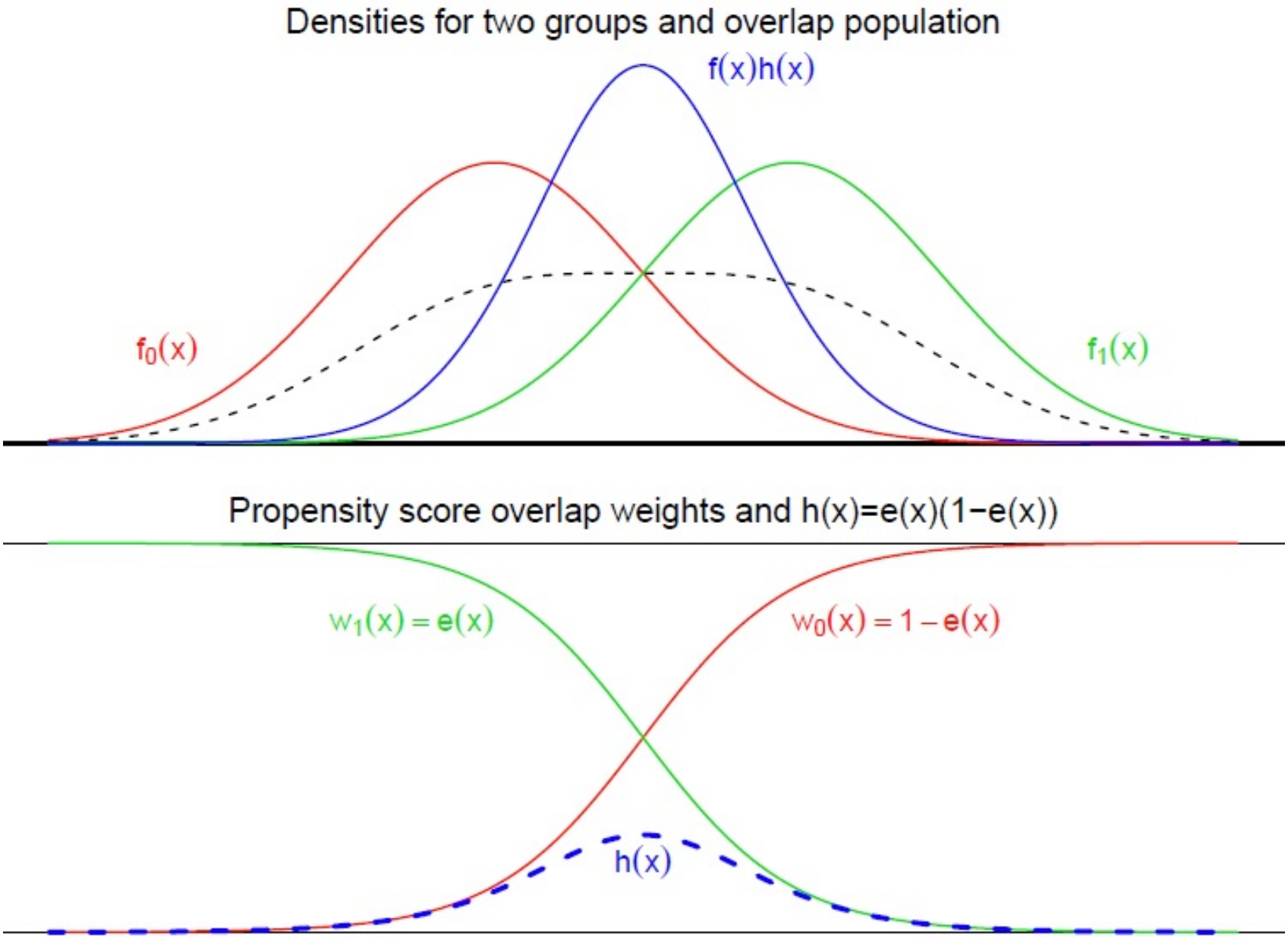
- As IPW, in practice one should always first normalize weights

Overlap Weights

- Review: $h(x) = e(x)(1 - e(x))$,

$$\begin{cases} w_1(x) \propto 1 - e(x), & \text{for } Z = 1, \\ w_0(x) \propto e(x), & \text{for } Z = 0. \end{cases}$$

- **QUESTION:** what value of the propensity score $e(x)$ gives highest $h(x)$?
  - **ANSWER:** $e(x) = 0.5$ yields $h(x) = 0.25$
  - The relative weighting of $f(x)$ by $h(x)$ is greatest at 0.5
- **Target population**: the units whose characteristics could appear with substantial probability in either treatment group (most overlap)
- $e(x)(1 - e(x))$: harmonic mean, also Gini index in classification tree literature


Densities for two groups and overlap population
f(x)h(x)
$f_0(x)$
$f_1(x)$
Propensity score overlap weights and h(x)=e(x)(1−e(x))
$w_1(x) = e(x)$
$w_0(x) = 1 - e(x)$
h(x)

Overlap Weights: Exact Balance

Theorem 3. *When the propensity scores are estimated by maximum likelihood under a logistic regression model,*

$$\text{logit}\{e(x_i)\} = \beta_0 + x_i'\beta,$$

the overlap weights lead to exact balance in the means of any included covariate between treatment and control groups:

$$\frac{\sum_{i=1}^n x_{ij} Z_i (1-\hat{e}_i)}{\sum_{i=1}^n Z_i (1-\hat{e}_i)} = \frac{\sum_{i=1}^n x_{ij} (1-Z_i)\hat{e}_i}{\sum_{i=1}^n (1-Z_i)\hat{e}_i}, \quad \text{for } j = 1, \ldots, p, \qquad (4)$$

where $\hat{e}_i = \{1 + \exp[-(\hat{\beta}_0 + x_i'\hat{\beta})]\}^{-1}$ *and* $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_j)$ *is the MLE for the regression coefficients.*

- Remark: the exact balance property applies to any included covariate and derived covariate, including high order terms and interaction terms of the covariates

Exact Balance: Outline of Proof

- The log likelihood function of the logistic PS model $\text{logit}\{e(X_i)\} = \beta_0 + X_i'\beta$ is:

$$L = \sum_{i=1}^{n} Z_i \log\{\hat{e}(x)\} + (1 - Z_i) \log\{1 - \hat{e}(x)\},$$

  where $\hat{e} = \left\{1 + \exp(-x_i'\beta)\right\}^{-1}$.

- To obtain the MLE of β's, one needs to solve for the score functions equating 0, which for each covariate k is:

$$S_k = \frac{\partial \log L}{\partial \beta_k} = \sum_{i=1}^{n} x_{ik}(Z_i - \hat{e}), \quad \text{for } k = 0, 1, \ldots, p, \tag{5}$$

  where $x_{0k} \equiv 1$.

- Equating to 0 and solving, the MLE $\hat{\beta}$ satisfies (exercise)

$$\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} \hat{e}_i, \quad \text{and} \quad \sum_{i=1}^{n} x_{ik} Z_i = \sum_{i=1}^{n} x_{ik} \hat{e}_i.$$

Overlap Weights: Exact Balance in Subgroups

(Yang et al. 2020)

Corollary 2. *If the postulated propensity score model includes any interaction term of a binary covariate, then the overlap weights lead to exact balance in the means in the subgroups defined by that binary covariate.*

Remarks:

- If the true PS model has interaction terms, then overlap weights using PS estimated from any model that nests the true model gives exact balance in the subgroups defined by the interaction terms

Exact Balance of Overlap Weights: More Remarks

- For binary treatment, the exact balance property only hold for the logistic link (the canonical link)
- Also holds if use a linear regression of Z on X to estimate $e(x)$: identity link is the canonical link
- Double-edge sword: The exact balance property holds regardless of the validity of the PS model – a peculiar math artifact of the canonical link of GLM
- Exact (mean) balance is a nice by-product of but not the main motivation behind overlap weights (minimize variance)
- In practice, one should always try to fit a good PS model even if this may break the exact balance

Overlap Weights: Statistical Advantages

- Minimum variance of the nonparametric estimator among all balancing weights (the homoscedasticity assumption is not crucial in practice)
- Exact balance for means of included covariates in logistic propensity score model
- Weights are bounded (unlike IPW)
- Avoids *ad hoc* truncating weights or eliminating cases: a continuous version of trimming

Overlap Weights: Adaptive

- Overlap weights are adaptive, ATO approximate
  - ATE: if treatment and control groups are nearly balanced in size and distribution (for $e(x) \approx 1/2$, $(1 - e(x), e(x)) \approx \left(\frac{.25}{e(x)}, \frac{.25}{1-e(x)}\right)$)
  - ATT: if propensity to treatment is always small (for $e(x) \approx 0$, $(1 - e(x), e(x)) \approx \left(1, \frac{e(x)}{1-e(x)}\right)$)
  - ATC: if propensity to control is small

Overlap Weights: Scientific Relevance

- Overlap weights focus on the (sub)population closest to the population in RCT
- Overlap weights put emphasis on internal validity
- The overlap population is of intrinsic substantive interest, for example
  - In medicine, patients in clinical equipoise
  - In policy, units whose treatment assignment would be most responsive to a policy shift as new information is obtained
- Overlap weighting mimics attributes of a randomized clinical trial. Useful in situations where large RCT is not immediately available, e.g. COVID-19 studies (Thomas et al. 2020b)
- Overlap weighting was first proposed by A Zaslavsky and applied in medical journals since 2001 (Schneider et al. 2001)

Characterize the Target Population

- An easy way to characterize and compare different target populations is to present a table of covariate means, both unadjusted and weighted (commonly known as Table 1 in medical papers)

	Unadjusted			Adjusted		
	Statin users N=348	No statins N=3008	Overall Unadjusted N=3356	IPTW Weighted N=3356	IPTW Sym Trim (0.10) N=1086	Overlap Weighted N=3356
Age – yr*	64 (57, 70)	57 (50, 65)	58 (51, 66)	58 (51, 66)	64 (58, 70)	63 (57, 69)
Female sex	44.3	54.3	53.2	52.1	41.8	44.2
Systolic Blood Pressure – mmHg	131 (119, 144)	125 (114, 138)	126 (114, 138)	126 (115, 139)	132 (120, 145)	131 (119, 144)
Diastolic Blood Pressure – mmHg	76 (69, 82)	75 (68, 81)	75 (69, 81)	75 (69, 81)	76 (69, 81)	76 (69, 81)
BMI – kg/m^2	29 (26, 32)	27 (25, 31)	28 (25, 31)	28 (25, 31)	29 (26, 32)	29 (26,32)
History of CVD						
Myocardial Infarction	16.4	3.0	4.4	5.9	13.6	12.7
Stroke	5.5	1.8	2.2	2.1	5.2	4.5
Peripheral Artery Dis.	7.5	2.2	2.7	2.7	6.9	5.9
Any ASCVD	35.3	8.9	11.6	13.4	32.8	29.4

Connection to Matching and Regression

- Matching: link “similar" cases in two samples, discard unmatched cases (bottom-up approach)
- Weighting: apply weights to entire samples, designed to create global balance (top-down approach)
- Intrinsic connection: Overlap weighting approaches many-to-many matching as the propensity score model becomes increasingly complex
- Why? The limit of a PS model is a saturated model with a fixed effect (dummy variable) for each design point (or small neighborhood, for continuous variables)

Connection to Matching and Regression

- Let the sample count for x_i in group $z = 0, 1$ be N_{zi}, the estimated PS from the saturated model is $\hat{e}(x_i) = N_{1i}/(N_{0i} + N_{1i})$
- The overlap weight for trt and con units is $N_{0i}/(N_{0i} + N_{1i})$ and $N_{1i}/(N_{0i} + N_{1i})$, respectively
- In many-to-many matching: at each design point x, each trt unit is matched to the N_{0i} controls at the same x, each control to N_{1i} trt units – exactly the same as the overlap weighting
- The total overlap weight for each design point x and hence of $\hat{\tau}(x_i)$ is $N_{0i}N_{1i}/(N_{0i} + N_{1i})$
- This is exactly the precision weight attached to $\bar{y}_{1i} - \bar{y}_{0i}$ in the fixed-effects for each design point OLS outcome model $Y_{zi} = \alpha_i + z\tau + \epsilon_{zi}$

Binary Outcomes: Estimands and Estimator

- Recall $\tau_1 = \Pr_g[Y(1) = 1] = \mathbb{E}_g[Y(1)] = \frac{\mathbb{E}_f[h(X)\mu_1(X)]}{\mathbb{E}_f(h(X))}$,
  $\tau_0 = \Pr_g[Y(0) = 1] = \mathbb{E}_g[Y(0)] = \frac{\mathbb{E}_f[h(X)\mu_0(X)]}{\mathbb{E}_f(h(X))}$

- $\tau^{\text{ATO}} = \tau_1 - \tau_0$ interpreted as the causal risk difference

- We define the causal risk ratio and odds ratio among the overlap population as

$$\tau_{\text{RR}} = \frac{\tau_1}{\tau_0}, \quad \tau_{\text{OR}} = \frac{\tau_1/(1-\tau_1)}{\tau_0/(1-\tau_0)}$$

- Estimate PS, and use

$$\hat{\tau}_1 = \frac{\sum_{i=1}^n Z_i Y_i (1-\hat{e}_i)}{\sum_{i=1}^n Z_i (1-\hat{e}_i)}, \quad \hat{\tau}_0 = \frac{\sum_{i=1}^n (1-Z_i) Y_i \hat{e}_i}{\sum_{i=1}^n (1-Z_i)\hat{e}_i}$$

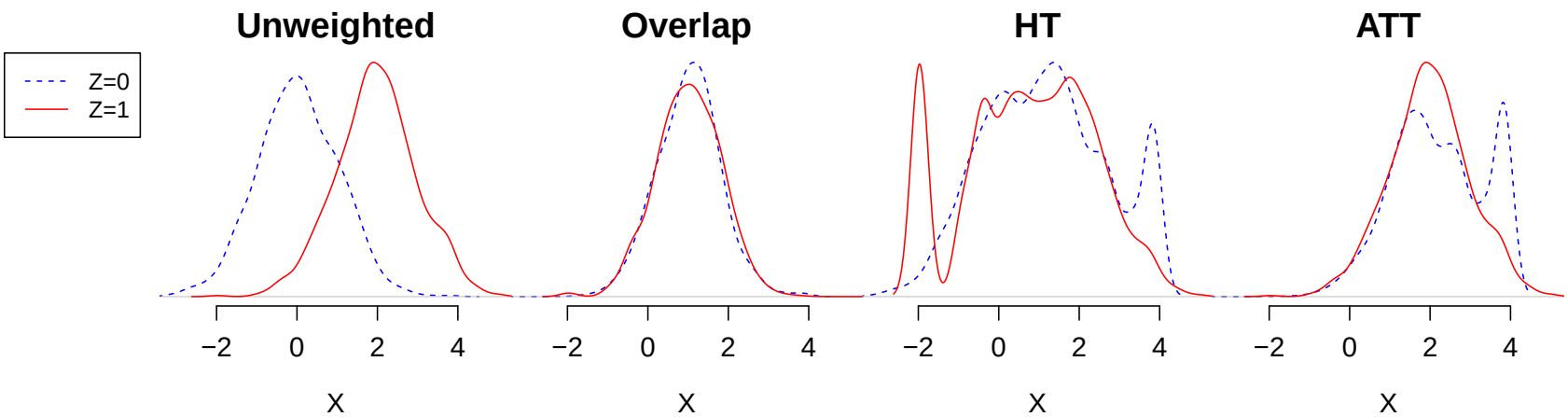
M-estimator of Variance

- A closed-form consistent variance estimator for τ_h using the OW (i.e. ATO estimand) when the PS is estimated from a logistic regression is given in Li, Thomas, Li (2019)
- Based on M-estimation, account for the uncertainty in estimating the propensity scores, good finite sample coverage
- Attractive in studies with large sample size
- If the outcome is binary and the estimand is causal odds ratio or relative risk, we can also derive corresponding closed-form variance estimator using the Delta method

A Simulated Example

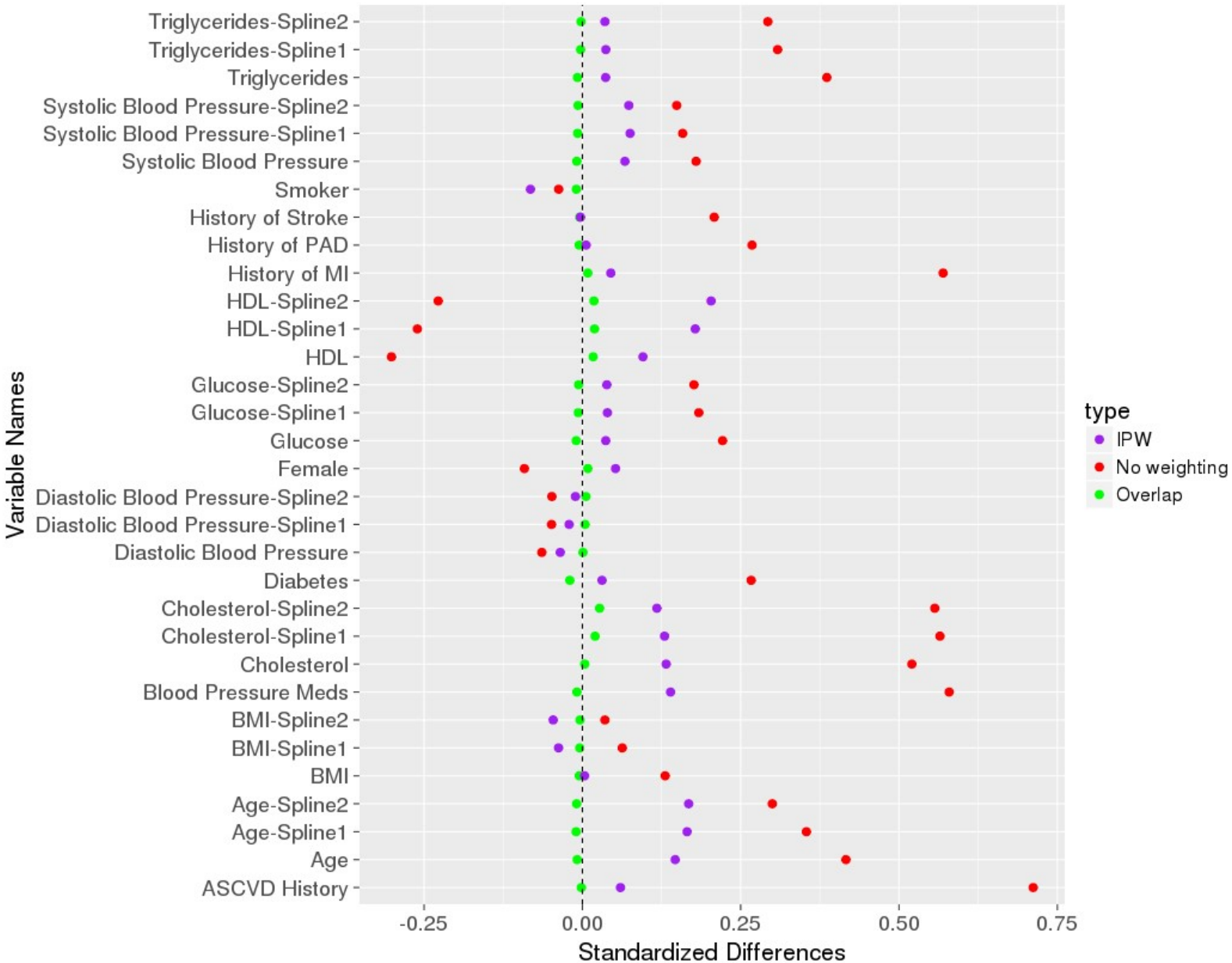
- Simulate $n_0 = n_1 = 1000$ units.
- A single covariate: $X_i \sim N(0,1) + 2Z_i$.
- Outcome model: $Y_i(z) \sim N(X_i, 1) + \tau z$, and $\tau = 1$.
- Use the nonparametric estimator $\hat{\tau}_h$ with different weights

Figure: Covariate distributions within each treatment group

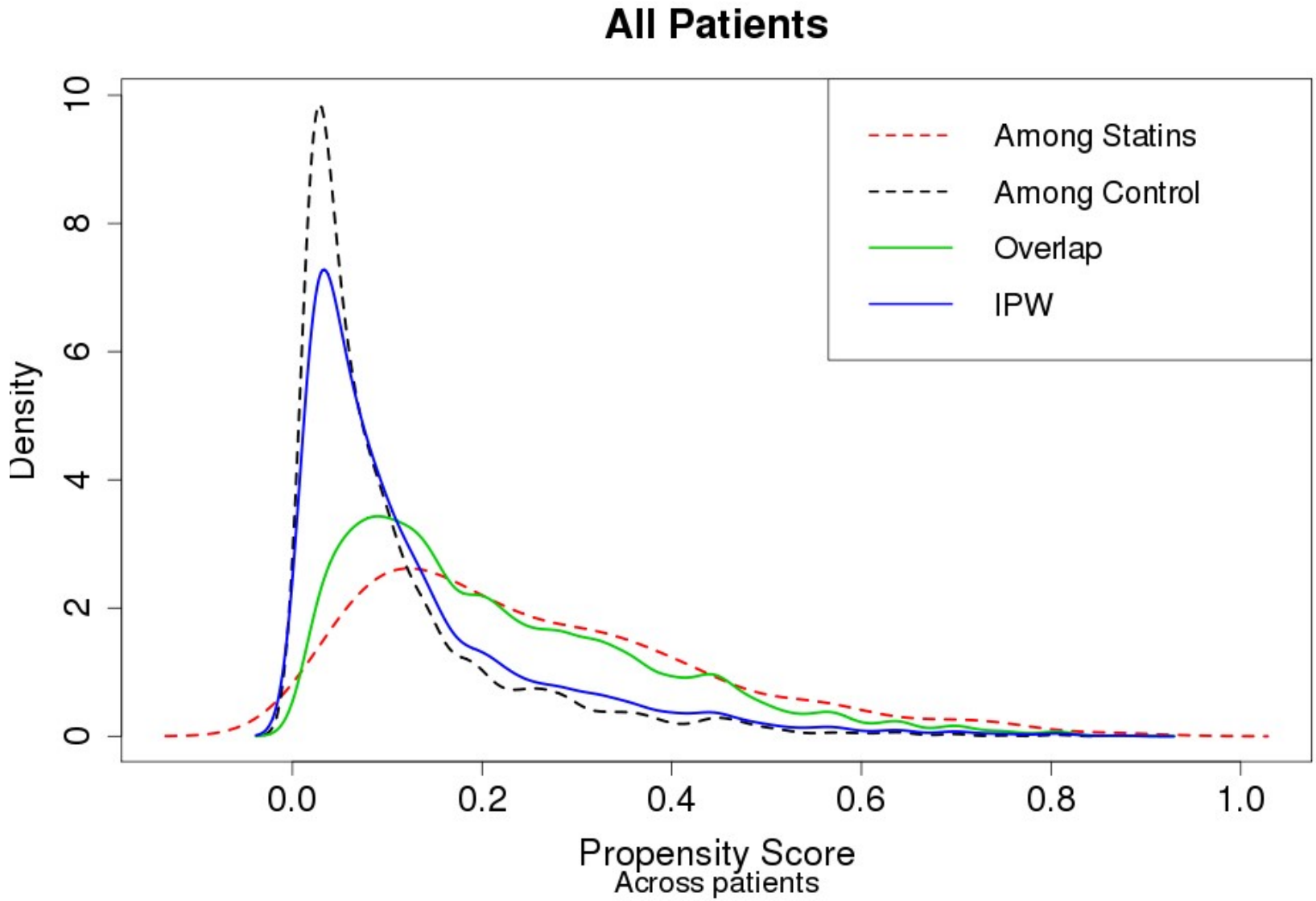


	Unweighted	Overlap	HT	ATT
$\hat{\tau}$	2.945	1.000	0.581	0.640
$SE(\hat{\tau})$	0.054	0.038	0.386	0.402

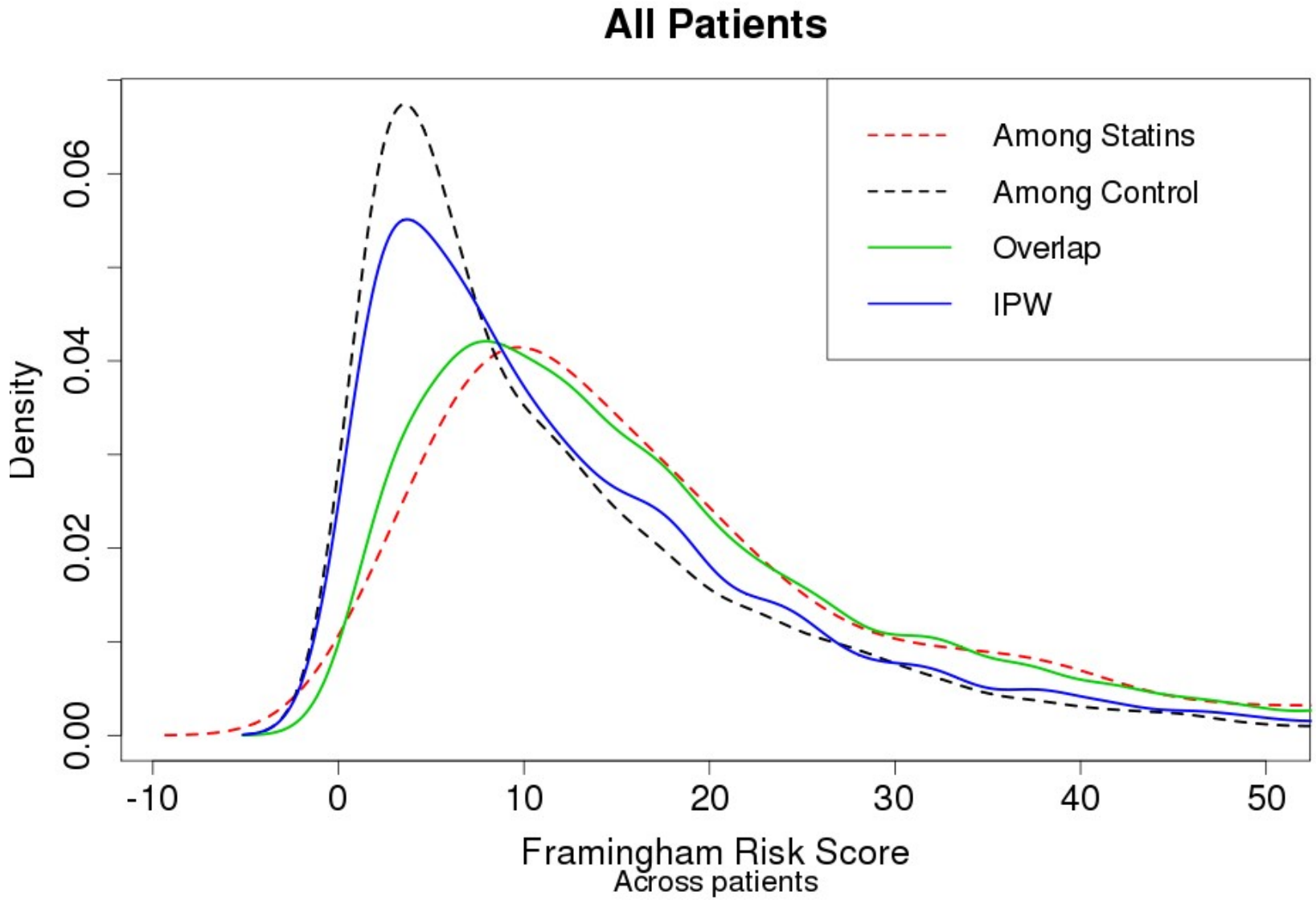
Framingham revisited



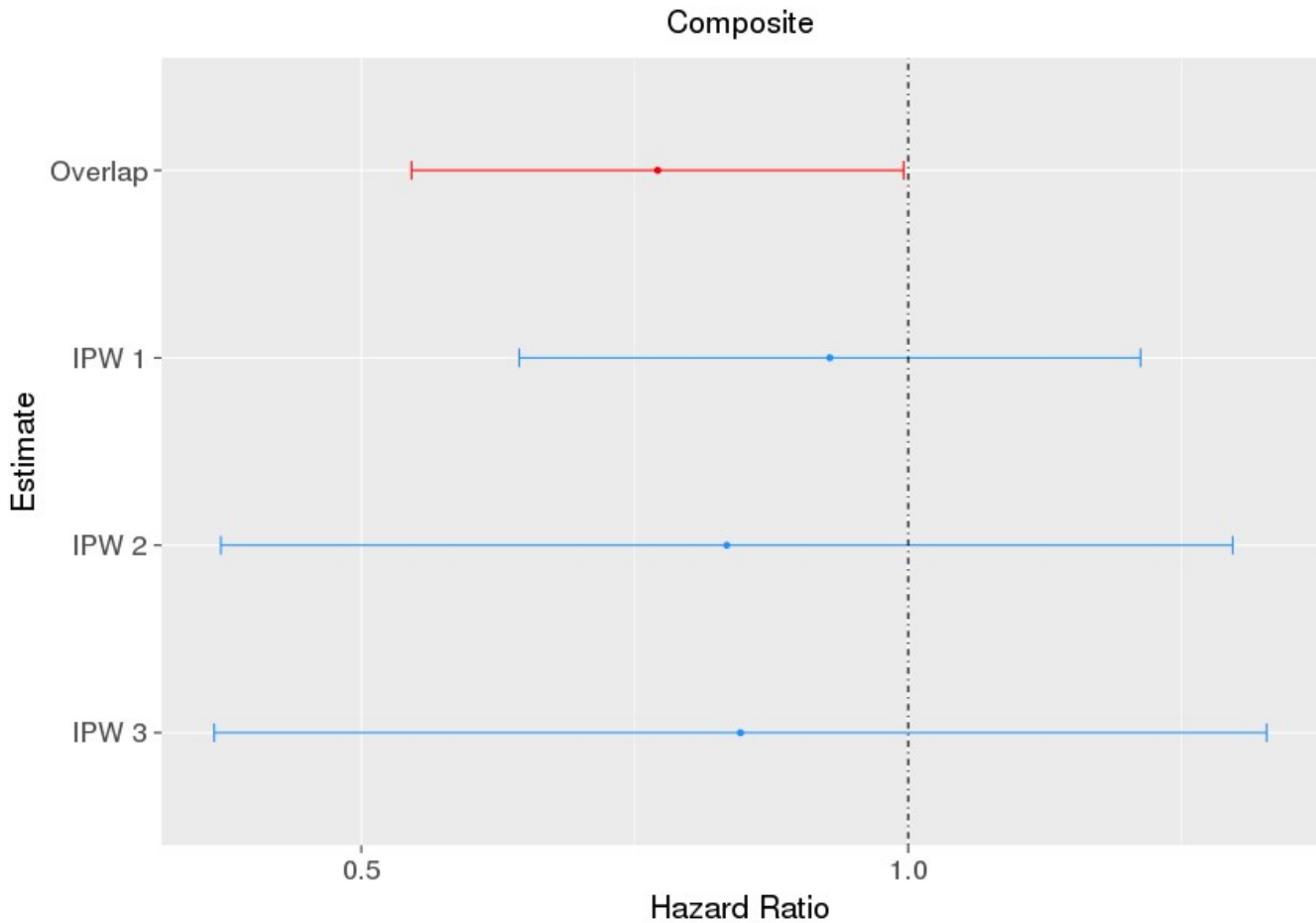
Weighted Distribution



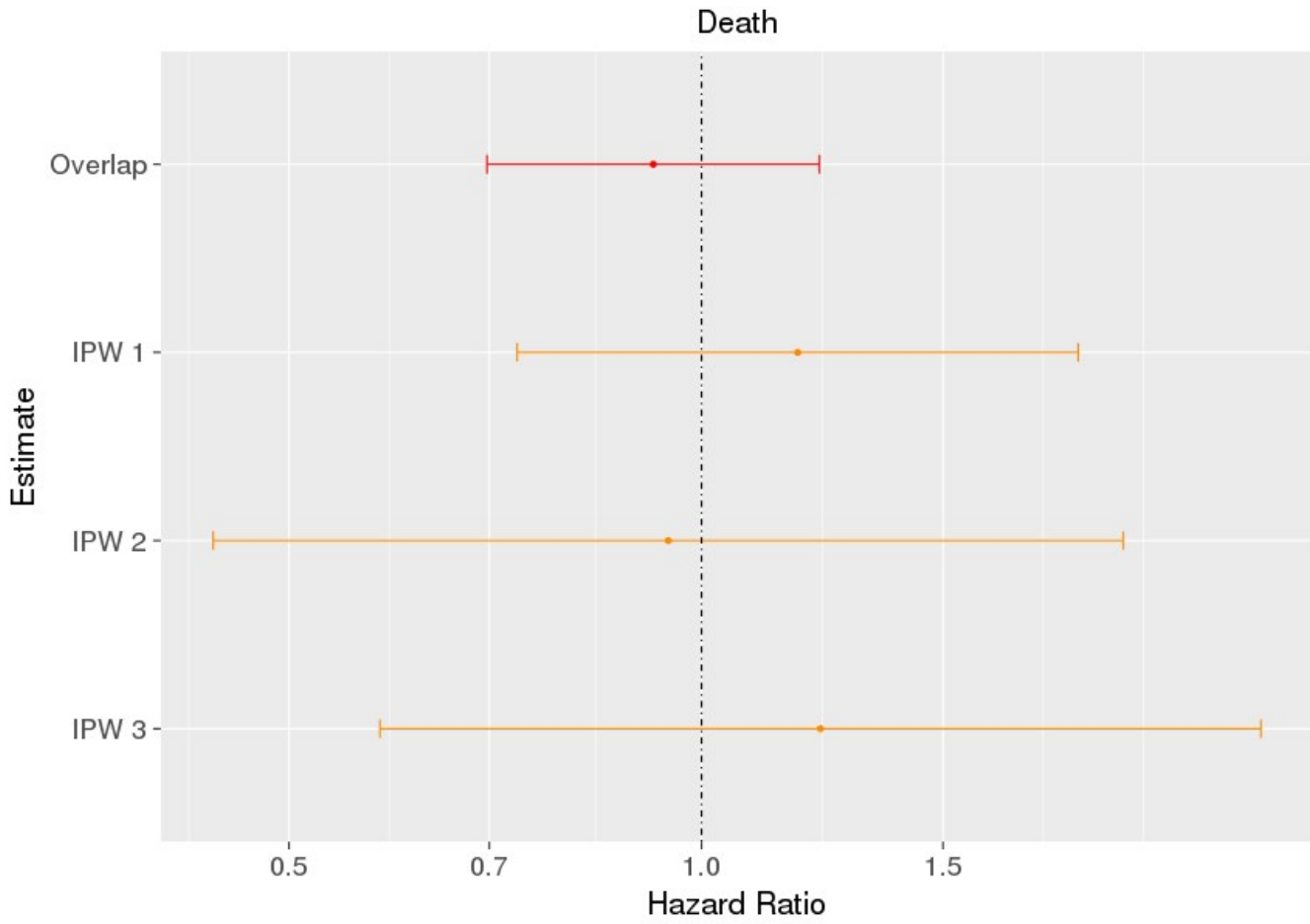
Weighted Distribution



Results: composite CV death



Results: all-cause mortality



PS Weighting for Covariate Adjustment in RCT

- PS weighting has been developed in the context of observational studies
- PS can also be used for covariate adjustment in randomized trials
- Chance imbalance is common in RCT; covariate adjustment improves precision
- ANCOVA with covariate-treatment interaction model (Yang and Tsiatis, 2001; Lin, 2013) is widely used, but potential for “fishing expedition", also lose efficiency if ANCOVA is misspecified
- In RCT, PS is known (often 1/2), how to do PS weighting?

PS Weighting for Covariate Adjustment in RCT

- PS weighting for covariate adjustment in RCT: procedure is no different from observational studies
  1. Specify a “working" PS model and obtain an estimated PS for each unit
  2. Use a PS weighting moment estimator based on the estimated PS as the covariate adjusted estimator
- Both IPW (Williamson et al., 2014; Shen et al., 2014) and OW (Zeng et al., 2020) works: asymptotically equivalent to the ANCOVA estimator, all belong to the same semiparametric family (Tsiatis et al. 2008)
- Key difference between RCT and observational studies: In RCT, PS is known and constant, so that the $h(x)$ for OW is constant, and thus IPW and OW corresponds to exact the same estimand - ATE

PS Weighting for Covariate Adjustment in RCT

- Finite sample performance: OW is more efficient than IPW, as good and sometimes even outperform an ANCOVA estimator with correct outcome model - the key is due to the exact balance property of OW
- PS weighting avoids outcome modeling, more objective and transparent for covariate adjustment
- This is an example of “estimated PS is better than true PS" (Rosenbaum, 1987)
- PS weighting methods with the family of balancing weights are implemented in the R package **PSweight** (available at CRAN) https://cran.r-project.org/web/packages/PSweight/index.html

Weighting vs. Balancing

- Balancing weights balance covariates in expectation, but do not balance in finite samples
- A recent stream of research proposed estimators that directly balance covariates, bypassing estimating PS
- Goal: find weights w_i such that the weighted average of each covariate is approximately balanced

$$\frac{1}{N_1}\sum_{i=1}^{N} w_i Z_i X_i = \frac{1}{N_1}\sum_{i=1}^{N} w_i (1 - Z_i) X_i$$

- General idea: minimize variation of weights (e.g. entropy, coef of variation) subject to a set of balance constraints
- Such weights are usually obtained via optimization algorithms

Weighting vs. Balancing

- Literature
  - Entropy balancing (Hainmueller, 2011)
  - Covariate Balancing Propensity Score (CBPS) (Imai and Ratkovic, 2014)
  - Stabilized balancing weights (Zubizarreta, 2015)
  - Inverse probability tilting (Graham, Pinto, Egel, 2012; 2016, JEBS)
  - Approximate residual balancing (Athey, Imbens, Wager, 2016)
- These estimators sacrifice some efficiency for finite-sample balance
- Open question: How much are these estimators better than a DR estimator with a flexible logistic model (e.g. semiparametric with power series or splines) for PS?

Covariate Balancing Propensity Score (CBPS)

Imai and Ratkovic (2014)

- In traditional PS analysis, one fits a model (e.g. logistic) for PS and then check the resulting balance and iteratively improve the fit
- Imai and Ratkovic (2014) proposed the Covariate Balancing Propensity Score (CBPS) method
- Core idea: fit a logistic PS model—i.e. optimize the likelihood function of the logistic model—subject to covariate balancing constraints

Covariate Balancing Propensity Score (CBPS)

Imai and Ratkovic (2014)

- Specifically, the logistic PS model is

$$\text{logit}\{e_\beta(X_i)\} = \beta_0 + X_i\beta', \tag{6}$$

and the balancing conditions are

$$\mathbb{E}\left\{\frac{Z_i f(X_i)}{e_\beta(X_i)}\right\} = \mathbb{E}\left\{\frac{(1-Z_i) f(X_i)}{1-e_\beta(X_i)}\right\}$$

where $f(X_i)$ be an M-dim vector-valued function of X_i

Covariate Balancing Propensity Score (CBPS)

Imai and Ratkovic (2014)

- In implementation, CBPS obtains the PS that optimizes the likelihood function of model (6) subject to (sample) constraints

$$\sum_{i=1}^{N}\left\{\frac{Z_i f(X_i)}{e_\beta(X_i)}\right\} = \sum_{i=1}^{n}\left\{\frac{(1-Z_i) f(X_i)}{1-e_\beta(X_i)}\right\}. \tag{7}$$

- When $f(X_i) = e'_\beta(X_i) = \partial e_\beta(X_i)/\partial\beta$, CBPS is exactly the MLE of the logistic model (6).
- Why? Look at the score functions of logistic model (6):

$$\sum_{i=1}^{n}\left\{\frac{Z_i e'_\beta(X_i)}{e_\beta(X_i)}\right\} = \sum_{i=1}^{n}\left\{\frac{(1-Z_i) e'_\beta(X_i)}{1-e_\beta(X_i)}\right\}. \tag{8}$$

- Balancing constraints are placed on covariates predictive of PS

Covariate Balancing Propensity Score (CBPS)

Imai and Ratkovic (2014)

- In the logistic model (6), we have $p + 1$ parameters and $p + 1$ constraints, the model is called “just-identified"
- We can also add more constraints, say on higher moments of covariates
- Then the PS model is over-identified: generally improve asymptotic efficiency at the cost of finite sample property – essentially a generalized method of moments (GMM) estimator (Newey and McFadden, 1994)

Covariate Balancing Propensity Score (CBPS)

Imai and Ratkovic (2014)

- Caveat: the PS model can be completely misspecified but the imposed moments are still balanced
- CBPS largely avoids extremely propensities, more robust and better balance
- Once CBPS is estimated, still apply to IPW to estimate ATE
- CBPS was extended to multiple and continuous treatments, longitudinal treatments (software and extension can be found on Kosuke Imai's webpage)

What is balancing all about?

- All the above balancing methods are designed to balance the functions (e.g. moments) of X that the analyst specifies
- A thorny question: what if some functions of X important (to the outcome model) are not specified?
- An example: if the true outcome model is $Y \sim X_1 + X_2 + X_1 X_2$, whereas we balance X_1, X_2, X_1^2, X_2^2 (and higher order moments) but not the interaction $X_1 X_2$, which may be severely imbalanced, then we will still end up with biased causal conclusion

What is balancing all about?

- Ultimately the outcome model (or generating mechanism) is the most important thing
- Elephant in the room: Not “how to balance" but “what to balance."
- Balancing is a (implicit) form of adjustment
- My general view: in observational studies, we need to ensure covariate balance to a certain extent (e.g., via PS), and then proceed with flexible outcome models

Aside 1: Propensity Score with Clustered Data

- Health care and policy data are often clustered, e.g. patients clustered in hospitals, clinics, insurance plans
- We focus on the case of individual assigned treatments with individual covariates and outcome information
- Main points (Li, Zaslavsky, Landrum, 2013):
  - Necessary to incorporating clustering information into at least one of, preferably both, stages of PS analysis
  - Various version of the unconfoundedness assumption: depend on the different conditioning sets, e.g. with or without cluster-specific fixed/random effects (open question)
  - Special care in standard error estimation

Propensity Score with Clustered Data

- Procedure of PS weighting with clustered data:
  - Stage 1: Estimate the propensity score using a fixed/random/mixed effects models with cluster-specific effects
  - Stage 2: Two methods
    - Method 1 (nonparametric): Calculate the cluster-specific treatment effect (using a selected weighting) within each cluster, and average these effects weighted by the total weights in each cluster
    - Methods 2 (augment by outcome model): augment weighting estimator by a multilevel (fixed/random/mixed effects) outcome model. Known as the double-robust estimator in the case of IPW
- Standard errors: Two methods
  - Bootstrap: resample the clusters (not individuals)
  - Augmented weighting: sandwich estimator (Lunceford and Davidian, 2003), taking into account the uncertainty of estimating PS
- Augmented estimators with both se methods are implemented in **PSweight** package for IPW, OW, MW

Aside 2: Propensity Score for Racial Disparities Analysis

- Causal studies: *no causation without manipulation* (Holland, 1986); a "cause" is at least hypothetically manipulable
- Race is obviously not manipulable
- Racial disparities in health care: not causal studies
- Question: Can propensity scores still be used for racial disparities? Hesitation and confusions among health service researchers
- Answer: Yes, with a few caveats.
- Key points:
  - The balancing property of PS does not involve potential outcomes nor imply causal interpretation
  - Racial disparities comparison requires weaker assumptions than causal comparisons

Controlled Descriptive Comparisons: Estimands

- Racial disparities comparison belongs to controlled descriptive comparisons:
  - Adjust for the difference in pre-specified covariates between two groups, rather than offer a causal interpretation
  - Other examples: different years, external vs. concurrent data (external controls)
- "Assignment": a nonmanipulable state defining membership in one of two groups or populations
- Conditional average controlled difference (ACD):

$$\tau(x) \equiv E(Y \mid Z = 1, X = x) - E(Y \mid Z = 0, X = x). \tag{9}$$

- Estimand: average controlled difference on a target population

$$\tau_h = E_g[\tau(X)] = \frac{E[h(X)\tau(X)]}{E[h(X)]}. \tag{10}$$

- τ_h: average net difference in outcome Y between two groups with the covariate dist adjusted to be the same as in the target population g

Causal vs. Controlled Descriptive Comparisons

Li and Li (2022), Observational Studies

- Assumptions for causal comparisons:
  - (A1) Stable Unit Treatment Value Assumption (SUTVA): (i) no interference, (ii)no different versions of the treatment
    - Necessary for defining two potential outcomes for each unit
  - (A2) Unconfoundedness: no unmeasured confounder
    - Necessary for connecting potential outcomes to the observed outcomes and interpreting the difference in the observed outcomes as a causal effect
  - (A3) Overlap or positivity: each unit has non-zero probability of being in either group.
- Assumptions for controlled descriptive comparisons
  - ACD does not involve potential outcomes; thus not require SUTVA or unconfoundedness
  - Overlap is still needed, for operational reasons (large variance o.w.)

Propensity Score for Racial Disparities Analysis

- Racial disparities in health care: a special case of controlled descriptive comparisons
- Propensity score analysis of racial disparities
  - Yes, PS can be used: the balancing property of PS does not involve potential outcomes nor imply causal interpretation
  - PS is a one-dimensional summary score that helps to achieve balance between covariates
- The concept of target population is also important in racial disparities. Different weights lead to different target population:
  - IPW: the overall population combining the two racial groups under comparison (an union of two sets)
  - OW: overlap population – the subpopulation with the most similar covariate distribution between the two racial groups (an intersection of two sets)
  - ATT weight: a population with the same covariate distribution as one pre-specified racial group (e.g. White or Black).

References

Hirano, K., Imbens, G. W., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), 1161-1189.

Robins, J. M., Hernan, M. A., Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology. 550-560

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., Sturmer, T. (2006). Variable selection for propensity score models. American journal of epidemiology, 163(12), 1149-1156.

Shortreed, S. M., Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. Biometrics, 73(4), 1111-1122.

Lunceford, J. K., Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19), 2937-2960.

Crump, RK, Hotz, VJ, Imbens, GW, and Mitnik, OA. (2006). Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika*, 96, 187-199.

Sturmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution-a simulation study. American journal of epidemiology, 172(7), 843-854.

References

Imai, K., Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 243-263.

Thomas, L. E., Li, F., Pencina, M. J. (2020). Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. Journal of American Medical Association. 323(23):2417-2418.

Schneider EC, Cleary PD, Zaslavsky AM, Epstein AM. (2001). Racial disparity in influenza vaccination: does managed care narrow the gap between African Americans and whites? *Journal of the American Medical Association*. 286(12):1455-60

Li F, and Li F. (2022). Using propensity scores for racial disparities. *Observational Studies*.

Li, F, Morgan, LK, and Zaslavsky, AM. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*. 113(521), 390-400.

Li, F, Thomas, LE, and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology*. 188(1), 250-257.

Li F, Zaslavsky AM, and Landrum MB. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19), 3373-3387.

References

Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in Medicine 2008; 27(23): 4658-4677.

Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. The American Statistician 2001; 55(4): 314-321.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. Annals of Applied Statistics, 7(1), 295-318.

Shen, C., Li, X., Li, L. (2014). Inverse probability weighting for covariate adjustment in randomized studies. Statistics in medicine, 33(4), 555-568.

Williamson, E. J., Forbes, A., White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. Statistics in medicine, 33(5), 721-737.

Zeng S, Li F, Wang R, Li, F. (2020). Propensity score weighting for covariate adjustment in randomized clinical trials. arXiv: 2004.10075

References

Rosenbaum, P. R. (1987). Model-based direct adjustment. Journal of the American Statistical Association, 82(398), 387-394.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 25-46.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511), 910-922.

Graham, B. S., de Xavier Pinto, C. C., Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. The Review of Economic Studies, 79(3), 1053-1079.

Athey, S., Imbens, G. W., Wager, S. (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. JRSS-B