Causal Inference Methods and Case Studies

STAT24630

Jingshu Wang

Lecture 14

Topic: Inverse probability weighting

- IPW
  - Using normalized weights
- Doubly robust estimator
- Bootstrap
- Textbook Chapters 17.8, 19.10.2

Motivation

- Matching methods can improve covariate balance
- Potential limitations of matching methods:
  - Inefficient: it may throw away many control units
  - Ineffective: it may not be able to balance covariates
  - Biased: not estimating the ATT if a lot of treated units are not matched

- Matching is a special case of weighting

$$\hat{\tau}_{\text{match}} = \frac{1}{N_t} \sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \frac{1}{|\mathcal{M}_i^c|} \sum_{i' \in \mathcal{M}_i^c} Y_{i'}^{\text{obs}} \right)$$

$$= \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i:W_i=1} \left(\frac{N_c}{N_t} \sum_{i':W_{i'}=1} \frac{1_{i \in \mathcal{M}_{i'}^c}}{|\mathcal{M}_{i'}^c|} \right) Y_i^{\text{obs}}$$

- Idea: weight each observation in the control group such that it looks like the treatment group

Inverse probability weighting (IPW)

- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right], \quad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]$$

  - Intuitively, unit that has a smaller $e(\boldsymbol{X}_i)$ has less chance to appear in the treatment group, so we should give it a higher weight (the less likely a subject is sampled, then the larger population it should represent)

$$\hat{\tau}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N} \frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1 - W_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)}$$

$$= \frac{1}{N}\sum_{i:W_i=1} \lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0} \lambda_i \cdot Y_i^{\text{obs}},$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i}} = \begin{cases} 1/(1 - e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

IPW for observational studies

- The propensity scores are estimated
- Estimate ATE and ATT
  - ATE

$$\widehat{\mathrm{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i Y_i^{\mathrm{obs}}}{\hat{e}(\boldsymbol{X}_i)} - \frac{(1-W_i)Y_i^{\mathrm{obs}}}{1-\hat{e}(\boldsymbol{X}_i)}\right\}$$

  - ATT

$$\widehat{\mathrm{ATT}} = \frac{1}{N_t}\sum_{i=1}^{N}\left\{W_i Y_i^{\mathrm{obs}} - \frac{\hat{e}(\boldsymbol{X}_i)(1-W_i)Y_i^{\mathrm{obs}}}{1-\hat{e}(\boldsymbol{X}_i)}\right\}$$

- For units that have identical propensity scores → difference-in-means estimator

Normalizing 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 (ω_i) by the sum of all weights in that group $\omega_i / \sum_{i':W_{i'}=w} \omega_{i'}$, for $w = 0,1$, i.e. the Hajek estimator:
- The new ATE estimator:

$$\widehat{\text{ATE}} = \frac{\sum_{i=1}^{N} W_i Y_i^{\text{obs}} / \hat{e}(\boldsymbol{X}_i)}{\sum_{i=1}^{N} W_i / \hat{e}(\boldsymbol{X}_i)} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i^{\text{obs}} / (1 - \hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^{N} (1 - W_i) / (1 - \hat{e}(\boldsymbol{X}_i))}$$

- The new ATT estimator:

$$\widehat{\text{ATT}} = \frac{1}{N_t} \sum_{i=1}^{N} W_i Y_i^{\text{obs}} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i^{\text{obs}} \hat{e}(\boldsymbol{X}_i) / (1 - \hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^{N} (1 - W_i) \hat{e}(\boldsymbol{X}_i) / (1 - \hat{e}(\boldsymbol{X}_i))}$$

- Using normalized weights, we can reduce variance and lead to more stable estimate (Hirano, Imbens, Ridder, 2003)

IPW advantages v.s. disadvantages

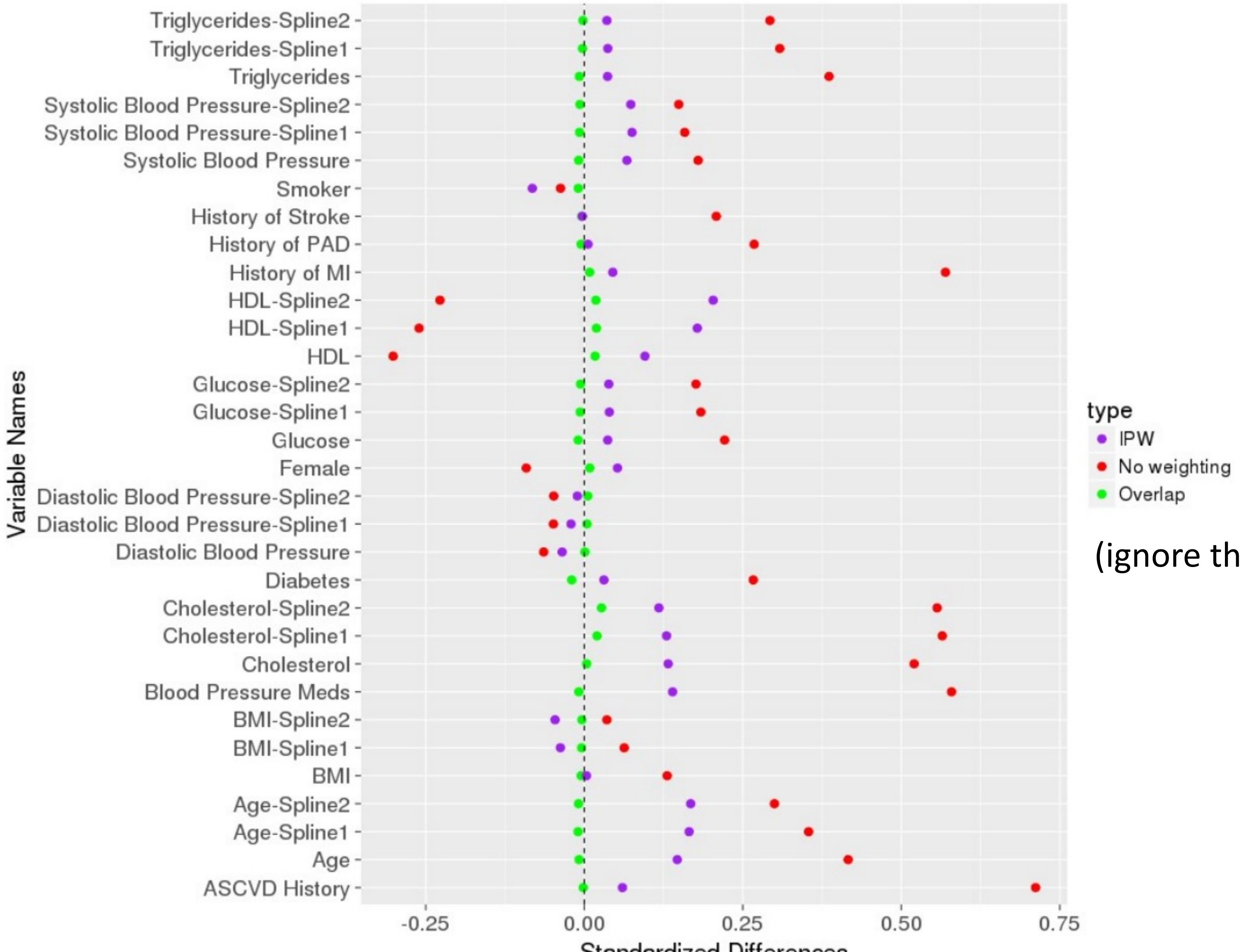
- Advantages
  - Simple, with theoretical foundation
  - Global balance
  - Use all data
  - Can be extended to more complex settings

- Disadvantages
  - More sensitive to misspecification of propensity scores then matching
  - Estimated propensity scores near 0 or 1 can yield extreme weights

Example: Framingham Heart Study

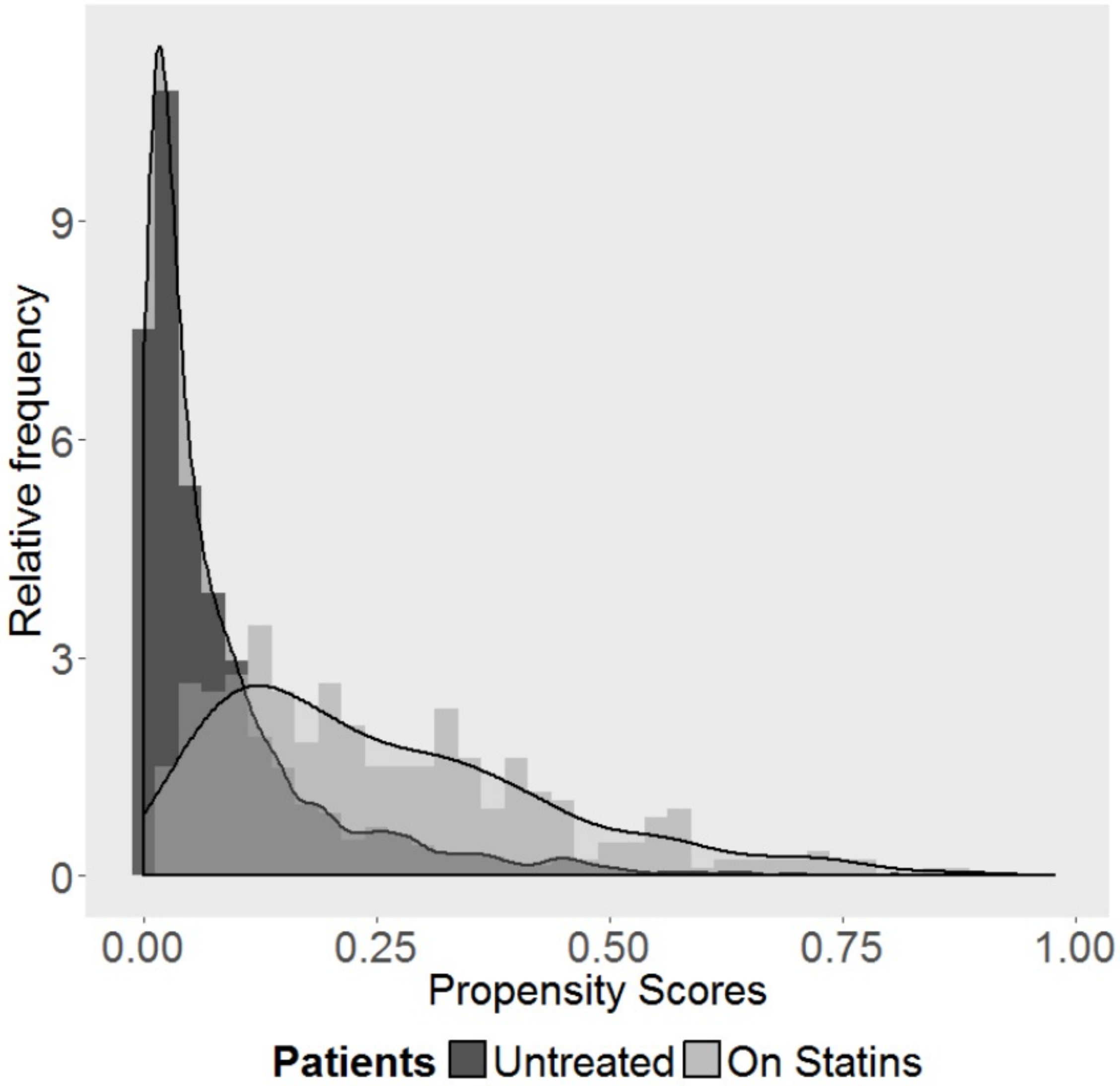
- Goal: evaluate the effect of statins on health outcomes
- Patients: cross-sectional population from the offspring cohort with a visit 6 (1995-1998)
- Treatment: statin use at visit 6 vs. no statin use
- Outcomes: CV(cardiovascular) death, myocardial infarction (MI), stroke
- Confounders: sex, age, 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)

Love plot for covariate balancing



(ignore the green dots)

Distribution of estimated propensity scores



- For treated units with $\hat{e}(\boldsymbol{X}_i)$ close to 0, then can greatly influence the IPW estimator value
- Will discuss trimming in later lectures

Doubly robust estimator

- Outcome regression relies on a correctly specified model for the (potential) outcomes depending on $\boldsymbol{X}_i$
- IPW / Matching relies on a correctly specified model for the propensity score

- Doubly robust estimator: provide a good estimate of the propensity score when either the outcome or the propensity score model is correct

- Define

$$f(1, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \frac{Y_i^{\text{obs}} 1_{W_i=1}}{\tilde{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=1} - \tilde{e}(\boldsymbol{X}_i)}{\tilde{e}(\boldsymbol{X}_i)} \tilde{\mu}_1(\boldsymbol{X}_i)$$

$$f(0, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \frac{Y_i^{\text{obs}} 1_{W_i=0}}{1 - \tilde{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=0} - (1 - \tilde{e}(\boldsymbol{X}_i))}{1 - \tilde{e}(\boldsymbol{X}_i)} \tilde{\mu}_0(\boldsymbol{X}_i)$$

  - If we correctly specify the propensity score model, then $\tilde{e}(\boldsymbol{X}_i) = e(\boldsymbol{X}_i)$
  - If we correctly specify the outcome model, then $\tilde{\mu}_w(\boldsymbol{X}_i) = \mu_w(\boldsymbol{X}_i)$

Doubly robust estimator

$$f(1, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \frac{Y_i^{\text{obs}} 1_{W_i=1}}{\tilde{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=1} - \tilde{e}(\boldsymbol{X}_i)}{\tilde{e}(\boldsymbol{X}_i)} \tilde{\mu}_1(\boldsymbol{X}_i)$$

$$f(0, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \frac{Y_i^{\text{obs}} 1_{W_i=0}}{1 - \tilde{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=0} - (1 - \tilde{e}(\boldsymbol{X}_i))}{1 - \tilde{e}(\boldsymbol{X}_i)} \tilde{\mu}_0(\boldsymbol{X}_i)$$

- $\tilde{e}(\boldsymbol{X}_i)$, $\tilde{\mu}_w(\boldsymbol{X}_i)$: our working models (model under our model assumption)
- $e(\boldsymbol{X}_i)$, $\mu_w(\boldsymbol{X}_i)$: true model that we don't know

- Double robust property

$$\mathbb{E}\left[f(1, \boldsymbol{X}_i, Y_i^{\text{obs}}) \mid \boldsymbol{X}_i\right] = \frac{(\mu_1(\boldsymbol{X}_i) - \tilde{\mu}_1(\boldsymbol{X}_i))\,(e(\boldsymbol{X}_i) - \tilde{e}(\boldsymbol{X}_i))}{\tilde{e}(\boldsymbol{X}_i)} + \mu_1(\boldsymbol{X}_i)$$

$$\mathbb{E}\left[f(0, \boldsymbol{X}_i, Y_i^{\text{obs}}) \mid \boldsymbol{X}_i\right] = \frac{(\mu_0(\boldsymbol{X}_i) - \tilde{\mu}_0(\boldsymbol{X}_i))\,(\tilde{e}(\boldsymbol{X}_i) - e(\boldsymbol{X}_i))}{1 - \tilde{e}(\boldsymbol{X}_i)} + \mu_0(\boldsymbol{X}_i)$$

  - If either the outcome or propensity score model is correct, we have

$$\mathbb{E}\left(f\left(w, \boldsymbol{X}_i, Y_i^{\text{obs}}\right)\right) = \mathbb{E}(Y_i(w) \mid \boldsymbol{X}_i)$$

Doubly robust estimator

$$f(1, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \underline{\frac{Y_i^{\text{obs}} 1_{W_i=1}}{\tilde{e}(\boldsymbol{X}_i)}} - \underline{\frac{1_{W_i=1} - \tilde{e}(\boldsymbol{X}_i)}{\tilde{e}(\boldsymbol{X}_i)} \tilde{\mu}_1(\boldsymbol{X}_i)}$$

IPW estimate of $\mathbb{E}(Y_i(1) | \boldsymbol{X}_i)$

Adjust for bias if the propensity score model is incorrect
(if PS model is correct, then this part has expectation 0)

An equivalent expression:

$$f(1, \boldsymbol{X}_i, Y_i^{\text{obs}}) = \underline{\tilde{\mu}_1(\boldsymbol{X}_i)} + \underline{\frac{1_{W_i=1}}{\tilde{e}(\boldsymbol{X}_i)} \left(Y_i(1) - \tilde{\mu}_1(\boldsymbol{X}_i)\right)}$$

Outcome regression estimate of $\mathbb{E}(Y_i(1) | \boldsymbol{X}_i)$

Adjust for bias if the outcome regression model is incorrect
(if PS model is correct, then this part has expectation 0)

The DR estimator:

$$\hat{\tau} = \frac{1}{N} \sum_i \left[\frac{Y_i^{\text{obs}} 1_{W_i=1}}{\widehat{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=1} - \widehat{e}(\boldsymbol{X}_i)}{\widehat{e}(\boldsymbol{X}_i)} \widehat{\mu}_1(\boldsymbol{X}_i) \right] - \frac{1}{N} \sum_i \left[\frac{Y_i^{\text{obs}} 1_{W_i=0}}{1 - \widehat{e}(\boldsymbol{X}_i)} - \frac{1_{W_i=0} - (1 - \widehat{e}(\boldsymbol{X}_i))}{1 - \widehat{e}(\boldsymbol{X}_i)} \widehat{\mu}_0(\boldsymbol{X}_i) \right]$$

A simulation study (Kang and Schafer. 2007. Statistical Science)

- The deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 4 covariates X_i^*: all are i.i.d. standard normal
  - Outcome model: linear model
  - Propensity score model: logistic model with linear predictors
  - Misspecification induced by measurement error:
    - $X_{i1} = \exp(X_{i1}^*/2)$
    - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
    - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
    - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$
- Weighting estimators to be evaluated:
  - HT: IPW in the original form
  - IPW: IPW with normalized weights
  - Weighted least squares regression with covariates
  - Doubly-robust estimator

Results: if the propensity score model is correct

		Bias		RMSE	
Sample size	Estimator	logit	True	logit	True
(1) Both models correct					
$n = 200$	HT	0.33	1.19	12.61	23.93
	IPW	−0.13	−0.13	3.98	5.03
	WLS	−0.04	−0.04	2.58	2.58
	DR	−0.04	−0.04	2.58	2.58
$n = 1000$	HT	0.01	−0.18	4.92	10.47
	IPW	0.01	−0.05	1.75	2.22
	WLS	0.01	0.01	1.14	1.14
	DR	0.01	0.01	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	−0.05	−0.14	14.39	24.28
	IPW	−0.13	−0.18	4.08	4.97
	WLS	0.04	0.04	2.51	2.51
	DR	0.04	0.04	2.51	2.51
$n = 1000$	HT	−0.02	0.29	4.85	10.62
	IPW	0.02	−0.03	1.75	2.27
	WLS	0.04	0.04	1.14	1.14
	DR	0.04	0.04	1.14	1.14

- Use the true propensity score is worse than using the estimated propensity score when the propensity score model is correct
- Normalizing weights can help a lot in reducing the variance

Results: if the propensity score model is incorrect

Sample size	Estimator	Bias logit	Bias True	RMSE logit	RMSE True
(3) Outcome model correct					
$n = 200$	HT	24.25	−0.18	194.58	23.24
	IPW	1.70	−0.26	9.75	4.93
	WLS	−2.29	0.41	4.03	3.31
	DR	−0.08	−0.10	2.67	2.58
$n = 1000$	HT	41.14	−0.23	238.14	10.42
	IPW	4.93	−0.02	11.44	2.21
	WLS	−2.94	0.20	3.29	1.47
	DR	0.02	0.01	1.89	1.13
(4) Both models incorrect					
$n = 200$	HT	30.32	−0.38	266.30	23.86
	IPW	1.93	−0.09	10.50	5.08
	WLS	−2.13	0.55	3.87	3.29
	DR	−7.46	0.37	50.30	3.74
$n = 1000$	HT	101.47	0.01	2371.18	10.53
	IPW	5.16	0.02	12.71	2.25
	WLS	−2.95	0.37	3.30	1.47
	DR	−48.66	0.08	1370.91	1.81

- Double robust estimator perform better when outcome model is correct but propensity score model is wrong
- Double robust estimator can perform worse when both models are wrong (maybe we should also normalize the weights in DR)

Variance of IPW estimator

- Researchers have shown that using the estimated propensity score asymptotically results in smaller variance of the IPW estimator (Hirano, Imbens and Ridder, 2003)
- 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
- In the R example, we show an approximation of the variance ignoring the uncertainty in estimating the propensity score by regression (not too bad, as the estimation of propensity score only involves pre-treatment covariates)

Bootstrap

	Parameter	Estimator	
F	$t(F)$	$\hat{t}$	
$F \to \hat{F}$	$t(\hat{F})$	$\hat{t}_1^*, \cdots, \hat{t}_B^*$	B bootstrap samples

- Nonparametric bootstrap:
  - Repeat B times: for each time b
    - sample N units with replacement (or resample the treated and controls separately)
    - Follow the whole procedure (starting from propensity score estimation to estimate the ATE/ATT using IPW)
    - Obtain an IPW estimator $\hat{\tau}_{IPW}^{(b)}$
  - Use the histogram of $\{\hat{\tau}_{IPW}^{(1)}, \cdots, \hat{\tau}_{IPW}^{(B)}\}$ as the approximated distribution of $\hat{\tau}_{IPW}$
    - The standard deviation of these estimates approximates the standard error of $\hat{\tau}_{IPW}$