STA 640 — Causal Inference

Chapter 3.3: Propensity Score Methods

Fan Li

Department of Statistical Science
Duke University

Propensity score

Definition (Rosenbaum and Rubin, 1983). The propensity score is defined as the conditional probability of receiving a treatment given pre-treatment covariates X:

$$e(X) = \Pr(Z = 1|X) = \mathbb{E}(Z|X), \tag{1}$$

where $X = (X_1, \ldots, X_p)$ is the collection of p covariates.

- Propensity score is a probability, analogous to a summary statistic (of the assignment mechanism)

Balancing property of propensity score

Property 1. The propensity score $e(X)$ balances the distribution of all X between the treatment groups:

$$Z \perp X \mid e(X).$$

Equivalently, $\Pr(Z_i = 1 \mid X_i, e(X_i)) = \Pr(Z_i = 1 \mid e(X_i))$.

- A balancing score $b(x)$ is a function of the covariates such that:

$$Z \perp X \mid b(X).$$

- Propensity score is a balancing score
- Rosenbaum and Rubin (1983) show that $e(X)$ is the coarsest balancing score: all balancing score is a function of $e(X)$

Remarks on the balancing property

1. If a subclass of units or a matched treatment-control pair is homogenous in $e(X)$, then the treatment and control units have the same distribution of X

2. If a subclass of units or a matched treatment-control pair is homogenous in both $e(X)$ and certain X, the other components of X within those refined class is also balanced – practical implication: estimating causal estimand in subpopulation, e.g. male or female group

3. The balancing property is a statement on the distribution of X, NOT on assignment mechanism or potential outcomes

Propensity score: Unconfoundedness

Property 2. If Z is unconfounded given X, then Z is unconfounded given $e(X)$, i.e.,

$$\{Y_i(1), Y_i(0)\} \perp Z_i \mid X_i \Longrightarrow \{Y_i(1), Y_i(0)\} \perp Z_i \mid e(X_i)$$

- Given a vector of covariates that ensure unconfoundedness, adjustment for differences in propensity scores removes all biases associated with differences in the covariates
- $e(X)$ can be viewed as a summary score of the observed covariates
- Causal inference can be drawn through stratification, matching, weighting, etc. using the scalar $e(X)$ instead of the high dimensional covariates.

Propensity score: remarks

- The propensity score balances the **observed** covariates, but does not generally balance **unobserved** covariates
- In most observational studies, the propensity score $e(X)$ is unknown and thus needs to be estimated
- There is a bias-variance tradeoff between modeling $e(X)$ and directly modeling the outcome $\Pr(Y(z) \mid X)$

Propensity score analysis of causal effects

Propensity score analysis (in observational studies) typically involves two stages:

Stage 1 Estimate the propensity score: by a logistic regression or machine learning methods, or covariate-balancing-type of propensity scores (CBPS)

Stage 2 Given the estimated propensity score, estimate the causal effects through one of these methods:

- Subclassification
- Weighting
- Matching
- Regression
- Mixed procedure of the above

Stage 1: Estimate propensity score

- The main purpose of estimating propensity score is to ensure overlap and balance of covariates between treatment groups, instead of “finding a perfect fit" of propensity score
- As long as the important covariates are balanced, model overfitting is usually not a concern (Rosenbaum, 1987)
- Essentially any balancing score (not necessarily propensity score) would be good enough for practical use
- A standard procedure for estimating prop score includes: initial fit, discard outliers, check covariate balance, re-fit if necessary

Stage 1: Estimate propensity score - overall flow

1. For binary treatments, most commonly via a logistic regression
2. Estimate propensity score by a logistic regression:

$$\text{logit } Pr(Z_i = 1 \mid X_i) = \beta X_i \tag{2}$$

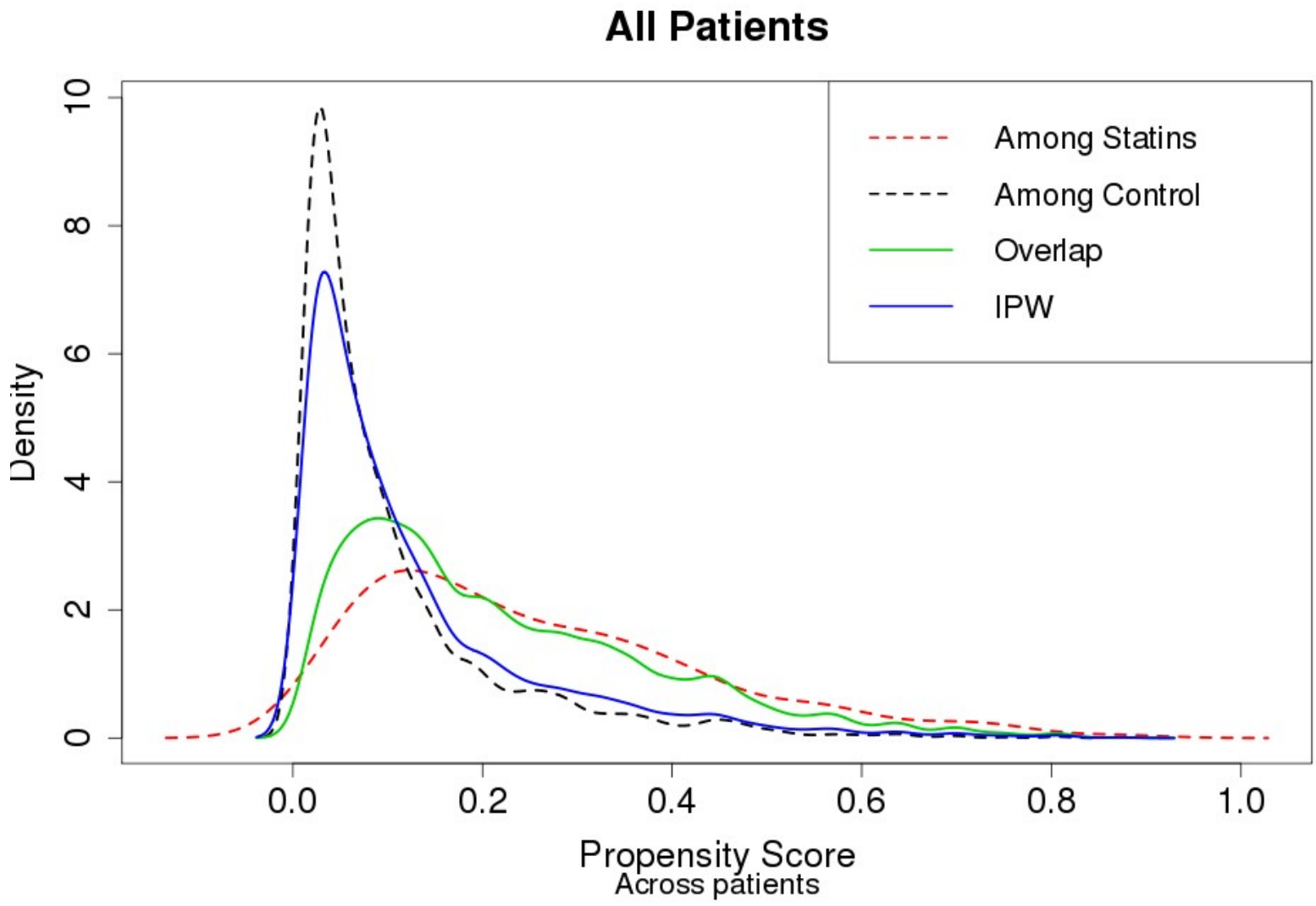
   by,e.g. a stepwise selection on the covariates and interactions) to get an initial estimate $e^0(X_i) = exp(\hat{\beta} X_i)/(1 + exp(\hat{\beta} X_i))$

3. Check overlap of propensity score between treatment groups. When the non-overlapping range is big, discard the observations with non-overlapping PS; otherwise, proceed to next step.
4. Assess balance given by initial model in (2), using matching, weighting or stratification (depends on estimation method in Stage 2)
5. If one or more covariates are seriously unbalanced: include some of their higher order terms, splines or/and interactions to re-fit the pscore model and repeat steps 2-3, until most covariates are balanced.

Stage 1: Estimate propensity score - check balance

- How to check balance depending on the estimation (of causal effects) method in Stage 2.
  - Stratification: check the balance (e.g. measured by ASD) of all important covariates within each stratum
  - Matching: check the balance of all important covariates in the matched sample
  - Weighting: check the balance of the weighted covariates between treatment and control groups.
- A useful way to visualize overlap/balance of the groups is to draw the histogram or density of the estimated PS by treatment and control groups

Density Plot of PS in Framingham Study



Stage 1: Estimate propensity score - stratification-based

1. After the initial propensity score $e^0(X_i)$ is estimated, check balance as follows
   - Create K blocks of $e^0(X_i)$ based on its quantiles (corresponding to K), $b_k, k = 1; \ldots K$
   - For every covariate, assess the balance within each of the block, e.g., by a t-test (large p-values means good balance)
   - For every covariate, also assess the overall balance, by performing an ANOVA of X on $Z \times b$. For true (or close to true) PS, we should expect no main effects of Z or interaction between Z and b due to the balance property
     1.1 Fit model 1: $X \sim b$ (b_i is block indictor of unit i);
     1.2 Fit model 2: $X \sim b + Z + Z \times b$;
     1.3 Perform an ANOVA on the predicted models 1 and 2 (a F-test), large p-values means good overall balance.

Case study: Swedish National March Cohort (NMC)

- NMC established in 1997: 300,000 Swedes participated in a Swedish Cancer Society national fund-raising event
- 43,880 individuals returned questionnaires including items on risk factors for cancer and cardiovascular disease (CVD): Physical activity (PA), age,sex, body mass index (BMI), etc
- These individuals were linked to CVD events from 1997 to 2004 via the Swedish patient registry
- Goal: study the causal effect of PA on CVD risk
- For this illustration, 10 covariates are selected

Case study: Swedish National March Cohort (NMC)

The model we used to estimate PS in the NMC data:

$$PA \sim male + age + age2 + bmi + sleep + smoking + smoke.former + alcohol + fitness + health + fittness \times age + bmi \times age + fitness \times health$$

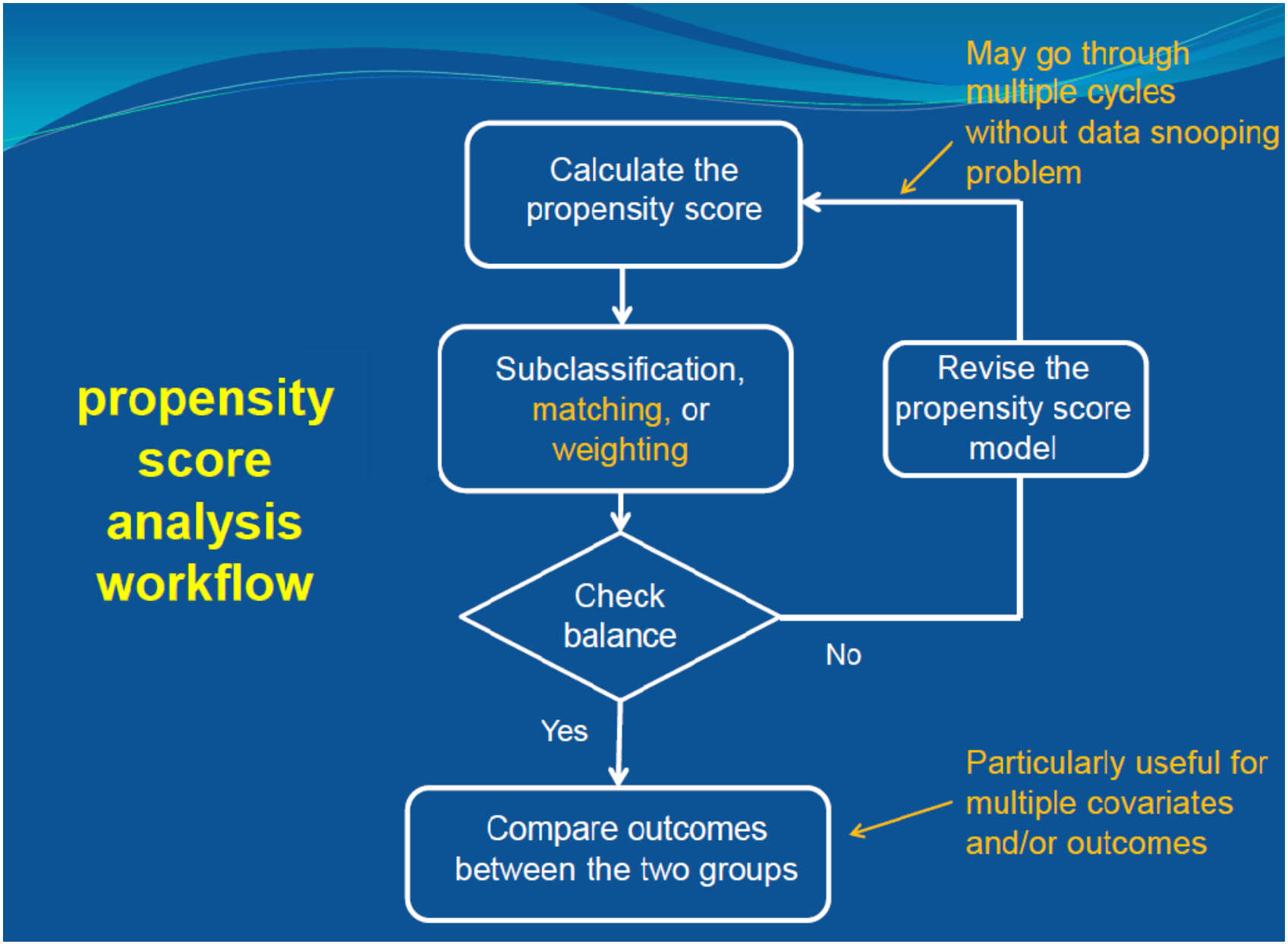
Table: P-values for balancing tests of the NMC data

X	block										Overall
	1	2	3	4	5	6	7	8	9	10	F-test
male	1	.97	.36	.36	.61	.24	.40	.17	.17	1	.51
age	.19	.67	.31	.86	.43	.35	.94	.11	.77	.00	.01
age2	.45	.97	.39	.52	.21	.72	.77	.19	.77	.04	.39
fitness	.02	.42	.45	.36	.48	.26	.82	.14	1	1	.02
bmi	.07	.00	.94	.76	.47	.29	.20	.05	.78	.86	.01
health	.86	.09	.73	.19	.94	.80	.95	.21	.77	.00	.18
smoking	.34	.69	.67	.68	.14	.18	.22	.83	.05	.64	.26
former smoker	.96	.47	.92	.65	.69	.87	.79	.87	.89	.72	.99
alcohol	.54	.05	.86	.15	.59	.42	.59	.84	.29	.95	.44
sleep	.41	.62	.29	.02	.85	.46	.81	.59	.96	.11	.73

Estimating propensity scores: flexible models

- Estimates of causal effects can be sensitive to mis-specification of propensity scores (Drake, 1993; Zhao, 2008)
- Flexible nonparametric or semiparametric models for the propensity score:
  - power series (Imbens, 2004)
  - splines (advocated by Rod Little et al.)
  - machine learning methods: ; CART, random forest etc. (Lee et al, 2010)
  - Ensemble learners: generalized boosted (McCaffrey et al., 2004; R package "twang"), super learners
  - Many more...
- In low dimensions, little difference, fancy models do NOT bring advantages (Alam et al. 2019)
- My experience: the old logistic regression with some interactions are usually good enough

Propensity Score Analysis Workflow



Stage 2: Stratification

- Recall Cochran's (1968) result of 5 subclasses of a single covariate removing 90% bias.
- Stratification using propensity score as the summary score should have approximately the same effects.
- Divide the subjects in to K strata by the corresponding quantiles of the estimated propensity score.
- Estimate ATE within each subclass of $e_k, (k = 1, \ldots, K)$ and then average by the block size:

$$\hat{\tau}^{\text{ATE}} = \sum_{k=1}^{K} \{(\bar{Y}_{k,1} - \bar{Y}_{k,0}\} \frac{N_{k,1}+N_{k,0}}{N},$$

  with $N_{k,1}, N_{k,0}$ being the numbers of units in class k under trt and control, respectively.
- Estimate ATT: weight the within-block ATE by the number of treated units $\hat{\tau}^{\text{ATT}} = \sum_{k=1}^{K} \hat{\tau}_k \cdot N_{k,1}/N_1$.

Propensity score stratification

- A variance estimator for $\hat{\tau}$ is

$$\mathbb{V}(\hat{\tau}) = \sum_{k=1}^{K} \{\mathbb{V}(\bar{Y}_{k,1}) + \mathbb{V}(\bar{Y}_{k,0})\}(\frac{N_{k,1}+N_{k,0}}{N})^2$$

- Bootstrap
- Case study: Rosenbaum and Rubin (1984)

Propensity score stratification: Remarks

- My experience: 5 blocks is usually not enough, consider higher number such as 10.
- Stratification is a coarsened version of matching (and weighting)
- Empirical results from real applications and situations: usually not as good as matching or weighting
- Good for cases with extreme outliers (smoothing): less sensitive, but also less efficient, and asymptotically biased (Lunceford and Davidian, 2004)
- Can be combined with regression: first estimate causal effects using regression within each block and then average the within-subclass estimates

Propensity score matching

- Special case of matching, the distance metric in the (estimated) propensity score
- 1-to-n closest neighbor matching is common when the control group is large compared to treatment group
- Pros: robust, matched pairs (within pair analysis), balance distributions in directions orthogonal to estimated PS
- Immensely popular, vast literature
- Sometimes, dimension reduction via the propensity score may be too drastic, recent methods advocate matching on the multivariate covariates directly
- Cons: inefficient, asymptotically biased (Abadie and Imbens, 2006)

Propensity score matching

- Interesting point by King and Nielsen (2019): *One shouldn't use propensity score for matching*.
- My view:
  - use matching as a pre-processing step only, one should combine matching with other methods
  - propensity score is an effective dimension reduction tool. Low dimension: little difference between methods; high dim: for matching, if not PS, then what?

Propensity score regression

- Remember the key propensity score property

$$\{Y_i(1), Y_i(0)\} \perp Z_i \mid X_i \Longrightarrow \{Y_i(1), Y_i(0)\} \perp Z_i \mid e(X_i)$$

- Idea: in a regression estimator, adjusting for $e(X)$ instead of the whole X; thus in regression models of $Y(z)$ use $e(X)$ as the single predictor
- Rubin (1985) argues the advantages are
  - Modeling $\Pr(Y(z)|e(X))$ is usually simpler than modeling $\Pr(Y(z)|X)$; effectively more data to estimate essential parameters due to the dimension reduction
  - When $\Pr(Y(z)|e(X))$ is correctly specified, Bayesian inference is well-calibrated in general and in all statements conditional on $e(X)$
- Further adjusting for the residuals in the regression of Y on $e(X)$ improves precision (Gutman and Rubin, 2015)

Propensity score regression

- Regression using PS as a single predictor $\Pr(Y(z)|\hat{e}(X))$ is not as efficient as the regression estimator adjusting for all covariates $\Pr(Y(z)|X)$ when $\Pr(Y(z)|X)$ is correctly specified (Hahn, 1998)
- The propensity score may be misspecified
- Simulations in Hade and Lu (2014) show: when the distributions of the estimated PS in the treated and control groups have different shapes but roughly the same support, regression on the estimated PS performs $\Pr(Y(z)|\hat{e}(X))$ well compared to estimator based on $\Pr(Y(z)|X)$
- Can be extended to more complex situations, e.g. sequential treatments (more later)

Propensity score regression

- Disadvantages of modeling $\Pr(Y(z)|\hat{e}(X))$:
  - Lose interpretation of the effects of individual covariates, e.g. age, sex
  - Reduction to the one-dimensional propensity score may be too drastic
- Idea: instead of using the estimated $\hat{e}(X)$ as the single predictor, use it as an additional predictor – model $\Pr(Y(z)|X, \hat{e}(X))$
- $\Pr(Y(z)|X, \hat{e}(X))$ gives both efficiency and robustness
- Empirical evidences (e.g. simulations) support
- Why it works? Continuous version of regression after stratification

Estimated versus True Propensity score

- A well known result on propensity score: using estimated propensity score usually has better efficiency and balance control than the true propensity score (Rosenbaum, 1987)
- Hirano, Imbens, Ridder (2003) provided theoretical basis using asymptotic arguments – asymptotic variance is smaller with estimated PS but not true PS. The proof is in the context of MLE.
- Intuition (Rosenbaum, 1987, pp 391): the same reason that covariate adjustment in RCT outperforms the unadjusted difference-in-means estimator – estimated PS corrects for chance imbalance in the sample, but true PS does not
- More later when we talk about PS weighting for covariate adjustment in RCT

References

Alam S., Moodie E. E. M., and Stephens D. A. (2019) Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Statistics in Medicine 38:1690-1702.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics, 1231-1236.

Gutman, R. and Rubin, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. Statistics in Medicine, 32(11): 1795-1814.

Hade, E. M., Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. Statistics in medicine, 33(1), 74-87.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica, 315-331.

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

G King and R Nielsen. (2019). Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27, 4

References

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.

McCaffrey, D. F., Ridgeway, G., Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods, 9(4), 403.

Rosenbaum P. (1987): Model-Based Direct Adjustment. Journal of the American Statistical Association, 82, 387-394.

Rosenbaum P, Rubin D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika; 70(1):41-55.

Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association, 79(387), 516-524.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. Bayesian statistics, 2, 463-472.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. Economics Letters, 98(3), 309-319.