Weighting

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- Under strong ignorability, we can use matching to account for the influence of confounders.

Review

- Under strong ignorability, we can use matching to account for the influence of confounders.
- We can directly match on the confounders (NN matching) or on the estimated propensity score (PS matching).

Review

- Under strong ignorability, we can use matching to account for the influence of confounders.
- We can directly match on the confounders (NN matching) or on the estimated propensity score (PS matching).
- NN matching is biased when there are more than one continuous confounders.

Review

- Under strong ignorability, we can use matching to account for the influence of confounders.
- We can directly match on the confounders (NN matching) or on the estimated propensity score (PS matching).
- NN matching is biased when there are more than one continuous confounders.
- But bias correction estimators exist and the bias is negligible under certain conditions.

Review

- Under strong ignorability, we can use matching to account for the influence of confounders.
- We can directly match on the confounders (NN matching) or on the estimated propensity score (PS matching).
- NN matching is biased when there are more than one continuous confounders.
- But bias correction estimators exist and the bias is negligible under certain conditions.
- Classical bootstrap does not work for NN matching.

Weighting

- Another choice is to estimate the propensity scores and use either the HT or the HA estimator.

Weighting

- Another choice is to estimate the propensity scores and use either the HT or the HA estimator.
- In observational studies, they are named as the inverse probability of treatment weighting (IPW) estimator and the stabilized IPW estimator.

Weighting

- Another choice is to estimate the propensity scores and use either the HT or the HA estimator.
- In observational studies, they are named as the inverse probability of treatment weighting (IPW) estimator and the stabilized IPW estimator.
- Since the propensity scores are estimated, neither estimator is unbiased.

Weighting

- Another choice is to estimate the propensity scores and use either the HT or the HA estimator.
- In observational studies, they are named as the inverse probability of treatment weighting (IPW) estimator and the stabilized IPW estimator.
- Since the propensity scores are estimated, neither estimator is unbiased.
- As long as the propensity score estimates are root-N consistent, both estimators are consistent and asymptotically normal.

Weighting

- Another choice is to estimate the propensity scores and use either the HT or the HA estimator.
- In observational studies, they are named as the inverse probability of treatment weighting (IPW) estimator and the stabilized IPW estimator.
- Since the propensity scores are estimated, neither estimator is unbiased.
- As long as the propensity score estimates are root-N consistent, both estimators are consistent and asymptotically normal.
- The estimation of the propensity scores introduces extra uncertainties into the ATE estimate.

Weighting: estimation

- Remember that the HT and HA estimators take the form of:

$$\hat{\tau}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{\hat{g}(\mathbf{X}_i)} - \frac{(1-D_i)Y_i}{1-\hat{g}(\mathbf{X}_i)}\right),$$

$$\hat{\tau}_{HA} = \frac{\sum_{i=1}^{N} D_i Y_i/\hat{g}(\mathbf{X}_i)}{\sum_{i=1}^{N} D_i/\hat{g}(\mathbf{X}_i)} - \frac{\sum_{i=1}^{N}(1-D_i)Y_i/(1-\hat{g}(\mathbf{X}_i))}{\sum_{i=1}^{N}(1-D_i)/(1-\hat{g}(\mathbf{X}_i))}.$$

Weighting: estimation

- Remember that the HT and HA estimators take the form of:

$$\hat{\tau}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{\hat{g}(\mathbf{X}_i)} - \frac{(1-D_i)Y_i}{1-\hat{g}(\mathbf{X}_i)}\right),$$

$$\hat{\tau}_{HA} = \frac{\sum_{i=1}^{N} D_i Y_i/\hat{g}(\mathbf{X}_i)}{\sum_{i=1}^{N} D_i/\hat{g}(\mathbf{X}_i)} - \frac{\sum_{i=1}^{N}(1-D_i)Y_i/(1-\hat{g}(\mathbf{X}_i))}{\sum_{i=1}^{N}(1-D_i)/(1-\hat{g}(\mathbf{X}_i))}.$$

- We can estimate the ATT or ATC using similar ideas:

$$\hat{\tau}_{HA,ATT} = \frac{\sum_{i=1}^{N} D_i Y_i}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N}(1-D_i)\hat{g}(\mathbf{X}_i)Y_i/(1-\hat{g}(\mathbf{X}_i))}{\sum_{i=1}^{N}(1-D_i)\hat{g}(\mathbf{X}_i)/(1-\hat{g}(\mathbf{X}_i))}.$$

Weighting: estimation

- To implement the Hajek estimator for the ATE, we first estimate the propensity score (using logistic regression) and obtain $\hat{g}(\mathbf{X}_i)$.

Weighting: estimation

- To implement the Hajek estimator for the ATE, we first estimate the propensity score (using logistic regression) and obtain $\hat{g}(\mathbf{X}_i)$.
- Next, we construct the weight,

$$W_i = \frac{D_i}{\hat{g}(\mathbf{X}_i)} + \frac{1 - D_i}{1 - \hat{g}(\mathbf{X}_i)}.$$

Weighting: estimation

- To implement the Hajek estimator for the ATE, we first estimate the propensity score (using logistic regression) and obtain $\hat{g}(\mathbf{X}_i)$.
- Next, we construct the weight,

$$W_i = \frac{D_i}{\hat{g}(\mathbf{X}_i)} + \frac{1 - D_i}{1 - \hat{g}(\mathbf{X}_i)}.$$

- Finally, we regress Y_i on D_i and weight each unit by W_i.

Weighting: estimation

- To implement the Hajek estimator for the ATE, we first estimate the propensity score (using logistic regression) and obtain $\hat{g}(\mathbf{X}_i)$.
- Next, we construct the weight,

$$W_i = \frac{D_i}{\hat{g}(\mathbf{X}_i)} + \frac{1 - D_i}{1 - \hat{g}(\mathbf{X}_i)}.$$

- Finally, we regress Y_i on D_i and weight each unit by W_i.
- For the Hajek estimator for the ATT,

$$W_i = D_i + \frac{(1 - D_i)\hat{g}(\mathbf{X}_i)}{1 - \hat{g}(\mathbf{X}_i)}.$$

Weighting: inference

- If $g(\mathbf{X}_i)$ is known, we can just use the HC2 variance estimator in regression.

Weighting: inference

- If $g(\mathbf{X}_i)$ is known, we can just use the HC2 variance estimator in regression.
- But $g(\mathbf{X}_i)$ is estimated.

Weighting: inference

- If $g(\mathbf{X}_i)$ is known, we can just use the HC2 variance estimator in regression.
- But $g(\mathbf{X}_i)$ is estimated.
- We can write

$$\hat{\tau} - \tau = \hat{\tau} - \hat{\tau}(g) + \hat{\tau}(g) - \tau,$$

where $\hat{\tau}(g)$ is the "oracle estimator:"

$$\hat{\tau}_{HT}(g) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{g(\mathbf{X}_i)} - \frac{(1-D_i)Y_i}{1-g(\mathbf{X}_i)}\right).$$

Weighting: inference

- If $g(\mathbf{X}_i)$ is known, we can just use the HC2 variance estimator in regression.
- But $g(\mathbf{X}_i)$ is estimated.
- We can write

$$\hat{\tau} - \tau = \hat{\tau} - \hat{\tau}(g) + \hat{\tau}(g) - \tau,$$

where $\hat{\tau}(g)$ is the "oracle estimator:"

$$\hat{\tau}_{HT}(g) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{g(\mathbf{X}_i)} - \frac{(1-D_i)Y_i}{1-g(\mathbf{X}_i)}\right).$$

- Then, we have

$$\begin{aligned} Var[\hat{\tau} - \tau] =& Var[\hat{\tau}(g) - \tau] + Var[\hat{\tau} - \hat{\tau}(g)] \\ & - 2Cov[\hat{\tau} - \hat{\tau}(g), \hat{\tau}(g) - \tau]. \end{aligned}$$

Weighting: inference

- If we conduct inference assuming $g(\mathbf{X}_i)$ is known, then the second and the third term in the expression are ignored.

Weighting: inference

- If we conduct inference assuming $g(\mathbf{X}_i)$ is known, then the second and the third term in the expression are ignored.
- However, we can prove that

$$Var[\hat{\tau} - \tau] \leq Var[\hat{\tau}(g) - \tau],$$

hence the variance will be conservative if we ignore the extra terms.

Weighting: inference

- If we conduct inference assuming $g(\mathbf{X}_i)$ is known, then the second and the third term in the expression are ignored.
- However, we can prove that

$$Var[\hat{\tau} - \tau] \leq Var[\hat{\tau}(g) - \tau],$$

  hence the variance will be conservative if we ignore the extra terms.
- For the HT estimator, if the logistic regression model is correctly specified and $m_{D_i}(\mathbf{X}_i)$ is smooth in $\mathbf{X}_i$, then the extra terms equal to

$$-E\left[\frac{(g(\mathbf{X}_i)m_0(\mathbf{X}_i) + (1 - g(\mathbf{X}_i))m_1(\mathbf{X}_i))^2}{g(\mathbf{X}_i)(1 - g(\mathbf{X}_i))}\right] \leq 0.$$

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!
- In a Bernoulli trial with $p_i = p$, the propensity score is p and the estimated propensity score is $\frac{\sum_{i=1}^{N} D_i}{N}$.

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!
- In a Bernoulli trial with $p_i = p$, the propensity score is p and the estimated propensity score is $\frac{\sum_{i=1}^{N} D_i}{N}$.
- Therefore, the IPW estimator using the true propensity scores is the Horvitz-Thompson estimator, while the one using the estimated propensity scores is the Hajek estimator.

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!
- In a Bernoulli trial with $p_i = p$, the propensity score is p and the estimated propensity score is $\frac{\sum_{i=1}^{N} D_i}{N}$.
- Therefore, the IPW estimator using the true propensity scores is the Horvitz-Thompson estimator, while the one using the estimated propensity scores is the Hajek estimator.
- We already know that the latter is more efficient!

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!
- In a Bernoulli trial with $p_i = p$, the propensity score is p and the estimated propensity score is $\frac{\sum_{i=1}^{N} D_i}{N}$.
- Therefore, the IPW estimator using the true propensity scores is the Horvitz-Thompson estimator, while the one using the estimated propensity scores is the Hajek estimator.
- We already know that the latter is more efficient!
- In observational studies, this conclusion was first proved by Hirano, Imbens, and Ridder (2003).

Weighting: inference

- Implication: Using the estimated propensity scores is more efficient than using the true propensity scores!
- In a Bernoulli trial with $p_i = p$, the propensity score is p and the estimated propensity score is $\frac{\sum_{i=1}^{N} D_i}{N}$.
- Therefore, the IPW estimator using the true propensity scores is the Horvitz-Thompson estimator, while the one using the estimated propensity scores is the Hajek estimator.
- We already know that the latter is more efficient!
- In observational studies, this conclusion was first proved by Hirano, Imbens, and Ridder (2003).
- Intuitively, estimating the propensity scores extracts more information from data.

Weighting: inference

- The IPW estimator using the estimated propensity scores can reach the efficiency bound derived by Hahn (1998).

Weighting: inference

- The IPW estimator using the estimated propensity scores can reach the efficiency bound derived by Hahn (1998).
- Let's define $\sigma_d^2(\mathbf{X}_i) = Var[Y_i(d)|\mathbf{X}_i]$, then the bound is

$$E\left[\frac{\sigma_1^2(\mathbf{X}_i)}{g(\mathbf{X}_i)} + \frac{\sigma_0^2(\mathbf{X}_i)}{1-g(\mathbf{X}_i)} + (m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \tau)^2\right].$$

Weighting: inference

- The IPW estimator using the estimated propensity scores can reach the efficiency bound derived by Hahn (1998).
- Let's define $\sigma_d^2(\mathbf{X}_i) = Var[Y_i(d)|\mathbf{X}_i]$, then the bound is

$$E\left[\frac{\sigma_1^2(\mathbf{X}_i)}{g(\mathbf{X}_i)} + \frac{\sigma_0^2(\mathbf{X}_i)}{1-g(\mathbf{X}_i)} + (m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \tau)^2\right].$$

- No estimator under strong ignorability can do better than this.

Weighting: inference

- The IPW estimator using the estimated propensity scores can reach the efficiency bound derived by Hahn (1998).
- Let's define $\sigma_d^2(\mathbf{X}_i) = Var[Y_i(d)|\mathbf{X}_i]$, then the bound is

$$E\left[\frac{\sigma_1^2(\mathbf{X}_i)}{g(\mathbf{X}_i)} + \frac{\sigma_0^2(\mathbf{X}_i)}{1-g(\mathbf{X}_i)} + (m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \tau)^2\right].$$

- No estimator under strong ignorability can do better than this.
- This conclusion is true under regularity conditions (low dimension and smoothness).

Weighting: inference

- The IPW estimator using the estimated propensity scores can reach the efficiency bound derived by Hahn (1998).
- Let's define $\sigma_d^2(\mathbf{X}_i) = Var[Y_i(d)|\mathbf{X}_i]$, then the bound is

$$E\left[\frac{\sigma_1^2(\mathbf{X}_i)}{g(\mathbf{X}_i)} + \frac{\sigma_0^2(\mathbf{X}_i)}{1 - g(\mathbf{X}_i)} + (m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \tau)^2\right].$$

- No estimator under strong ignorability can do better than this.
- This conclusion is true under regularity conditions (low dimension and smoothness).
- The high-dimensional case is analyzed by Su et al. (2023).

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.
- Then, the HT estimator will have a huge variance.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.
- Then, the HT estimator will have a huge variance.
- The HA estimator performs better in this case.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.
- Then, the HT estimator will have a huge variance.
- The HA estimator performs better in this case.
- One choice to trim units whose propensity score takes extreme values (Yang and Ding 2018; Ma and Wang 2020).

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.
- Then, the HT estimator will have a huge variance.
- The HA estimator performs better in this case.
- One choice to trim units whose propensity score takes extreme values (Yang and Ding 2018; Ma and Wang 2020).
- It alters the estimand and causes bias.

Weighting: pros and cons

- Weighting does not require bias correction or drop any units.
- But we need to have accurate predictions for the propensity score.
- There is a trade-off between convergence rate and accuracy.
- In practice, the estimated propensity score can be very close to 0 or 1.
- It is caused by the failure of positivity.
- Then, the HT estimator will have a huge variance.
- The HA estimator performs better in this case.
- One choice to trim units whose propensity score takes extreme values (Yang and Ding 2018; Ma and Wang 2020).
- It alters the estimand and causes bias.
- Another choice is to use the covariate balancing propensity scores (Imai and Ratkovic 2014).

Covariate balancing propensity scores

- We have proved that the propensity score is a balance score:

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

Covariate balancing propensity scores

- We have proved that the propensity score is a balance score:

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

- We can exploit this property to improve the accuracy of our estimation.

Covariate balancing propensity scores

- We have proved that the propensity score is a balance score:

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

- We can exploit this property to improve the accuracy of our estimation.
- For any function of the covariates, $f(\mathbf{X}_i)$, we should have:

$$E\left[\frac{D_i f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i)f(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] = 0.$$

Covariate balancing propensity scores

- We have proved that the propensity score is a balance score:

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

- We can exploit this property to improve the accuracy of our estimation.
- For any function of the covariates, $f(\mathbf{X}_i)$, we should have:

$$E\left[\frac{D_i f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i)f(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] = 0.$$

- In finite sample, we should expect

$$\sum_{i=1}^{N}\left[\frac{D_i f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i)f(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] \approx 0.$$

Covariate balancing propensity scores

- We have proved that the propensity score is a balance score:

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

- We can exploit this property to improve the accuracy of our estimation.
- For any function of the covariates, $f(\mathbf{X}_i)$, we should have:

$$E\left[\frac{D_i f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i) f(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] = 0.$$

- In finite sample, we should expect

$$\sum_{i=1}^{N}\left[\frac{D_i f(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i) f(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] \approx 0.$$

- We can set $f(\mathbf{X}_i)$ to be each of the covariates or their higher order terms.

Covariate balancing propensity scores

- Remember that we often estimate the propensity score via the logistic model:

$$g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i\beta}}{1 + e^{\mathbf{X}_i\beta}}.$$

Covariate balancing propensity scores

- Remember that we often estimate the propensity score via the logistic model:
$$g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i\beta}}{1+e^{\mathbf{X}_i\beta}}.$$
- The first order condition is
$$\sum_{i=1}^{N}\left[\frac{D_i g'(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1-D_i)g'(\mathbf{X}_i)}{1-g(\mathbf{X}_i)}\right] = 0.$$

Covariate balancing propensity scores

- Remember that we often estimate the propensity score via the logistic model:

$$g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i \beta}}{1 + e^{\mathbf{X}_i \beta}}.$$

- The first order condition is

$$\sum_{i=1}^{N} \left[\frac{D_i g'(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1 - D_i) g'(\mathbf{X}_i)}{1 - g(\mathbf{X}_i)} \right] = 0.$$

- Therefore, the same balance condition holds for $g'(\mathbf{X}_i)$ as well.

Covariate balancing propensity scores

- Remember that we often estimate the propensity score via the logistic model:
$$g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i\beta}}{1 + e^{\mathbf{X}_i\beta}}.$$
- The first order condition is
$$\sum_{i=1}^{N} \left[\frac{D_i g'(\mathbf{X}_i)}{g(\mathbf{X}_i)} - \frac{(1 - D_i) g'(\mathbf{X}_i)}{1 - g(\mathbf{X}_i)} \right] = 0.$$
- Therefore, the same balance condition holds for $g'(\mathbf{X}_i)$ as well.
- We can combine all these balance conditions to estimate propensity scores more precisely.

Covariate balancing propensity scores

- Suppose we know $g(\mathbf{X}_i)$, then the probability for us to observe $\mathcal{D} = (D_1, D_2, \ldots, D_N)$ is

$$L = \prod_{i=1}^{N} g(\mathbf{X}_i)^{D_i}(1 - g(\mathbf{X}_i))^{1-D_i}.$$

Covariate balancing propensity scores

- Suppose we know $g(\mathbf{X}_i)$, then the probability for us to observe $\mathcal{D} = (D_1, D_2, \ldots, D_N)$ is

$$L = \prod_{i=1}^{N} g(\mathbf{X}_i)^{D_i}(1 - g(\mathbf{X}_i))^{1-D_i}.$$

- We find $\hat{\beta}$ such that

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta} L \\ &= \arg\max_{\beta} \log L \\ &= \arg\max_{\beta} \sum_{i=1}^{N} [D_i \log(g(\mathbf{X}_i)) + (1 - D_i)\log(1 - g(\mathbf{X}_i))]\end{aligned}$$

Covariate balancing propensity scores

- Suppose we know $g(\mathbf{X}_i)$, then the probability for us to observe $\mathcal{D} = (D_1, D_2, \ldots, D_N)$ is

$$L = \prod_{i=1}^{N} g(\mathbf{X}_i)^{D_i}(1 - g(\mathbf{X}_i))^{1-D_i}.$$

- We find $\hat{\beta}$ such that

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} L \\
&= \arg\max_{\beta} \log L \\
&= \arg\max_{\beta} \sum_{i=1}^{N} \left[D_i \log(g(\mathbf{X}_i)) + (1 - D_i)\log(1 - g(\mathbf{X}_i))\right]
\end{aligned}$$

- $\hat{\beta}$ can be found via the first order condition.

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^{N} = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^{N}$, all the balance conditions are satisfied.

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^N = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^N$, all the balance conditions are satisfied.
- In logistic regression, we have only one balance condition.

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^N = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^N$, all the balance conditions are satisfied.
- In logistic regression, we have only one balance condition.
- Estimation could be done by the Generalized Method of Moments (Hansen 1982).

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^N = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^N$, all the balance conditions are satisfied.
- In logistic regression, we have only one balance condition.
- Estimation could be done by the Generalized Method of Moments (Hansen 1982).
- Suppose we have K balance conditions: $\Psi(\mathbf{p}) = (\Psi_1(\mathbf{p}), \Psi_2(\mathbf{p}), \ldots, \Psi_K(\mathbf{p}))$, where $\mathbf{p} = (p_1, p_2, \ldots, p_N)$.

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^N = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^N$, all the balance conditions are satisfied.
- In logistic regression, we have only one balance condition.
- Estimation could be done by the Generalized Method of Moments (Hansen 1982).
- Suppose we have K balance conditions: $\Psi(\mathbf{p}) = (\Psi_1(\mathbf{p}), \Psi_2(\mathbf{p}), \ldots, \Psi_K(\mathbf{p}))$, where $\mathbf{p} = (p_1, p_2, \ldots, p_N)$.
- We try the following problem

$$\hat{\beta} = \arg\min_{\beta} \hat{E}[\Psi]' \widehat{Var}^{-1}[\Psi] \hat{E}[\Psi]$$

Covariate balancing propensity scores

- We try to find an estimate $\hat{\beta}$ such that with $\{\hat{p}_i\}_{i=1}^N = \{\hat{g}(\mathbf{X}_i)\}_{i=1}^N$, all the balance conditions are satisfied.
- In logistic regression, we have only one balance condition.
- Estimation could be done by the Generalized Method of Moments (Hansen 1982).
- Suppose we have K balance conditions: $\Psi(\mathbf{p}) = (\Psi_1(\mathbf{p}), \Psi_2(\mathbf{p}), \ldots, \Psi_K(\mathbf{p}))$, where $\mathbf{p} = (p_1, p_2, \ldots, p_N)$.
- We try the following problem

$$\hat{\beta} = \arg\min_{\beta} \hat{E}[\Psi]' \widehat{Var}^{-1}[\Psi] \hat{E}[\Psi]$$

- We then rely on these $\{\hat{p}_i\}_{i=1}^N$ to construct the IPW estimators.

Covariate balancing propensity scores

- CBPS can handle continuous treatment variables (Fong, Hazlett, and Imai 2018).

Covariate balancing propensity scores

- CBPS can handle continuous treatment variables (Fong, Hazlett, and Imai 2018).
- We find weights that are orthogonal to $\mathbf{X}$, D, and their interaction

$$\sum_{i=1}^{N} p_i \mathbf{X}_i = 0, \sum_{i=1}^{N} p_i D_i = 0$$

$$\sum_{i=1}^{N} p_i (\mathbf{X}_i * D_i) = 0, \sum_{i=1}^{N} p_i = N$$

Covariate balancing propensity scores

- CBPS can handle continuous treatment variables (Fong, Hazlett, and Imai 2018).
- We find weights that are orthogonal to $\mathbf{X}$, D, and their interaction

$$\sum_{i=1}^{N} p_i \mathbf{X}_i = 0, \sum_{i=1}^{N} p_i D_i = 0$$

$$\sum_{i=1}^{N} p_i (\mathbf{X}_i * D_i) = 0, \sum_{i=1}^{N} p_i = N$$

- The properties of CBPS are derived in Fan et al. (2016).

Covariate balancing propensity scores

- CBPS can handle continuous treatment variables (Fong, Hazlett, and Imai 2018).
- We find weights that are orthogonal to $\mathbf{X}$, D, and their interaction

$$\sum_{i=1}^{N} p_i \mathbf{X}_i = 0, \sum_{i=1}^{N} p_i D_i = 0$$

$$\sum_{i=1}^{N} p_i (\mathbf{X}_i * D_i) = 0, \sum_{i=1}^{N} p_i = N$$

- The properties of CBPS are derived in Fan et al. (2016).
- CBPS forces the propensity scores to balance the covariates, hence the estimates are less likely to be extreme.

Weighting: application

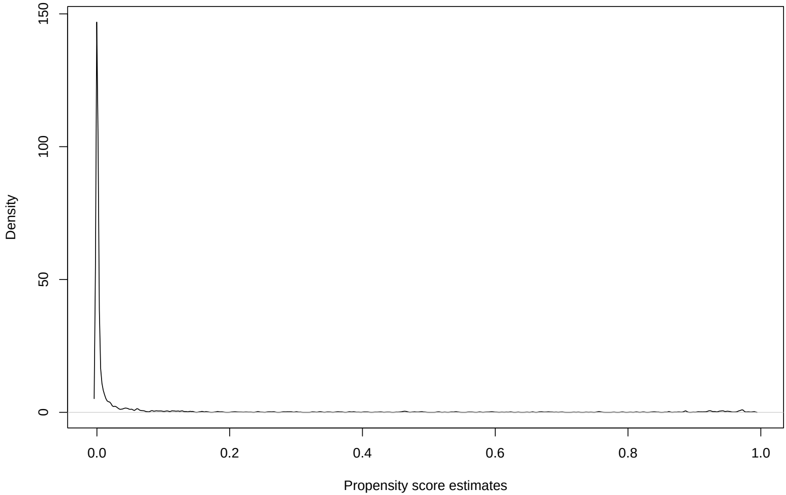
```
## The OLS estimate is 1794.343

## The SE of OLS estimate is 670.9967

## The Lin regression estimate is 1583.468

## The SE of Lin regression estimate is 678.0574
```

Weighting: application



Weighting: application

```
## The IPW ATT estimate is 2796.213

## The SE of IPW ATT estimate is 862.6273

##               mean.Tr  mean.Co   sdiff T pval
## age            25.816   23.812  28.012  0.002
## education      10.346   10.286   2.977  0.772
## black           0.843    0.818   6.853  0.483
## hispanic        0.059    0.120 -25.665  0.021
## married         0.189    0.093  24.449  0.005
## nodegree        0.708    0.716  -1.698  0.858
## re74         2095.574 1434.631  13.526  0.130
## re75         1532.056 1344.515   5.826  0.511
## u74             0.708    0.812 -22.795  0.012
## u75             0.600    0.413  38.134  0.000
```

Weighting: application

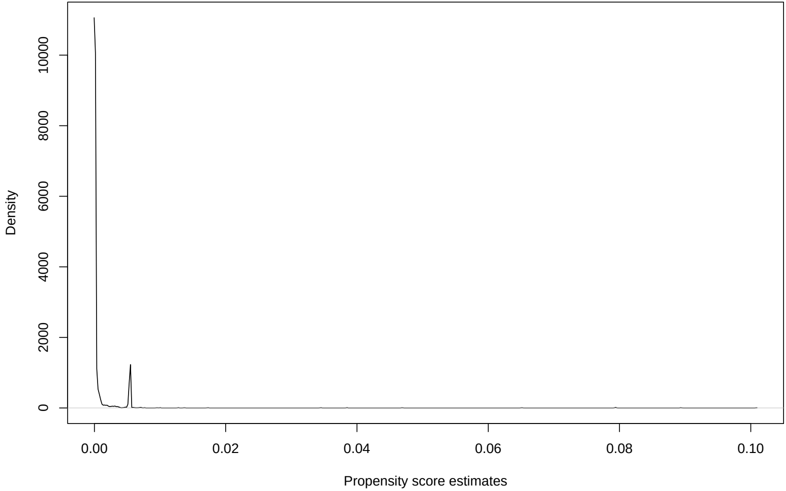
```
## The IPW ATT estimate is 1767.74

## The SE of IPW ATT estimate is 1116.721

##              mean.Tr  mean.Co   sdiff T pval
## age           27.021   25.382  23.171  0.053
## education     10.479   10.844 -17.496  0.136
## black          0.826    0.886 -15.679  0.151
## hispanic       0.069    0.025  17.252  0.080
## married        0.188    0.131  14.349  0.194
## nodegree       0.674    0.691  -3.689  0.753
## re74        1900.917 2186.548  -7.085  0.580
## re75        1204.959 1682.989 -23.763  0.107
## u74            0.681    0.688  -1.492  0.899
## u75            0.604    0.635  -6.182  0.598
```

Weighting: application

```
## [1] "Finding ATT with T=1 as the treatment.  Set ATT=2 t
```



Weighting: application

```
## The IPW ATT estimate is 2437.704

## The SE of IPW ATT estimate is 896.333
```

References I

Fan, Jianqing, Kosuke Imai, Han Liu, Yang Ning, Xiaolin Yang, et al. 2016. "Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach." *URL: Https://Imai. Fas. Harvard. Edu/Research/CBPStheory. Html*.

Fong, Christian, Chad Hazlett, and Kosuke Imai. 2018. "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements." *The Annals of Applied Statistics* 12 (1): 156–77.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*, 315–31.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica: Journal of the Econometric Society*, 1029–54.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89.

References II

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 243–63.

Ma, Xinwei, and Jingshen Wang. 2020. "Robust Inference Using Inverse Probability Weighting." *Journal of the American Statistical Association*, 1–10.

Su, Fangzhou, Wenlong Mou, Peng Ding, and Martin J Wainwright. 2023. "When Is the Estimated Propensity Score Better? High-Dimensional Analysis and Bias Correction." *arXiv Preprint arXiv:2303.17102*.

Yang, Shu, and Peng Ding. 2018. "Asymptotic Inference of Causal Effects with Observational Studies Trimmed by the Estimated Propensity Scores." *Biometrika* 105 (2): 487–93.