Average Causal Effect in Observational Studies II

Xinzhou Guo

HKUST

(Credited to Zhichao Jiang)

February 26, 2024

$$E[Y(1)] = E\left[\frac{Z_i Y_i}{e(X_i)}\right] = E[\mu_1(X_i)]$$

↓ ↓

Method IPW OR.

↓

Hybrid Doubly Robust

Other estimation strategies

- Other use of propensity score in estimating ACE besides stratification, IPW, doubly robust estimator
  1. Can we use propensity in regression? – reweight and covariate
  2. How is regression with propensity related to the others?
- Calibration methods – without modelling the propensity score
- Matching – distance or propensity

Regression weighted by inverse of propensity score

- Hajek estimator

$$\widehat{\text{ACE}}^{\text{hajek}} = \frac{\sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}(\mathbf{X}_i)}}{\sum_{i=1}^{n} \frac{Z_i}{\hat{e}(\mathbf{X}_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-Z_i) Y_i}{1-\hat{e}(\mathbf{X}_i)}}{\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}(\mathbf{X}_i)}}$$

- Weighted least squares estimation – why and what is the interpretation?

$$\left(\hat{\beta}_0^{\text{wls}}, \hat{\beta}_1^{\text{wls}}\right) = \underset{\beta_0, \beta_1}{\text{argmin}} \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 Z_i\right)^2$$

$$w_i = \frac{Z_i}{\hat{e}\left(\mathbf{X}_i\right)} + \frac{1 - Z_i}{1 - \hat{e}\left(\mathbf{X}_i\right)} = \begin{cases} \frac{Z_i}{\hat{e}(\mathbf{X}_i)} & \text{if } Z_i = 1 \\ \frac{1-Z_i}{1-\hat{e}(\mathbf{X}_i)} & \text{if } Z_i = 0 \end{cases}$$

- Variance estimation: bootstrap but how?

$$\hat{\beta} = \begin{pmatrix} \sum w_i & \sum w_i z_i \\ \sum w_i z_i & \sum w_i z_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum w_i y_i \\ \sum w_i z_i y_i \end{pmatrix}$$

$$= \frac{1}{\sum \frac{z_i}{e_i} \sum \frac{1-z_i}{1-e_i}} \begin{pmatrix} \cdots & \cdots \\ -\sum \frac{z_i}{e_i} & \frac{z_i}{e_i} + \frac{1-z_i}{1-e_i} \end{pmatrix}$$

$$y = \alpha + \beta x + \gamma z + \delta xz + \varepsilon$$

Correctly specify: $\hat{\beta}_{OLS} \rightarrow \beta = AC\hat{\varepsilon}$

$\hat{\gamma}_{WLS} \rightarrow \gamma = AC\hat{\varepsilon}$

Incorrect: $\hat{\gamma}_{WLS} \rightarrow \gamma$

Weighted regression with covariates

- In randomized experiments, we run $\mathrm{lm}\,(Y_i \sim 1 + Z_i + X_i + Z_i X_i)$ to improve efficiency
- In observational studies, can we run $\mathrm{lm}\,(Y_i \sim 1 + Z_i + X_i + Z_i X_i)$ with weights – what's the role of weight?
  - equivalent to two regressions: $\mathbb{E}\,(Y_i \mid Z_i = z, X_i) = \alpha_z + \gamma_z^\top X_i$ – does models have to be correctly specified?
  - ACE estimator: $\widehat{\alpha}_1^{\mathrm{wls}} - \widehat{\alpha}_0^{\mathrm{wls}}$
  - How about the propensity is misspecified?

① What is estimator after reweighting

② bad of property

Weighted regression with covariates

- Consider a doubly robust estimator based on linear models

$$\mathbb{E}\left(Y_i \mid Z_i = z, X_i\right) = \alpha_z + \gamma_z^\top X_i$$

$$\widehat{\mu}_{1,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i\left\{Y_i - \mu_1\left(\mathbf{X}_i, \widehat{\alpha}_1^{\mathrm{wls}}, \widehat{\gamma}_1^{\mathrm{wls}}\right)\right\}}{e\left(\mathbf{X}_i, \widehat{\beta}\right)} + \mu_1\left(\mathbf{X}_i, \widehat{\alpha}_1^{\mathrm{wls}}, \widehat{\gamma}_1^{\mathrm{wls}}\right)\right]$$

$$\widehat{\mu}_{0,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-Z_i)\left\{Y_i - \mu_0\left(\mathbf{X}_i, \widehat{\alpha}_0^{\mathrm{wls}}, \widehat{\gamma}_0^{\mathrm{wls}}\right)\right\}}{1 - e\left(\mathbf{X}_i, \widehat{\beta}\right)} + \mu_0\left(\mathbf{X}_i, \widehat{\alpha}_0^{\mathrm{wls}}, \widehat{\gamma}_0^{\mathrm{wls}}\right)\right]$$

$$\widehat{\mathrm{ACE}}_{\mathrm{DR}} = \widehat{\mu}_{1,\mathrm{DR}} - \widehat{\mu}_{0,\mathrm{DR}}$$

- From the property of weighted least squares

$$\widehat{\mathrm{ACE}}_{\mathrm{DR}} = \widehat{\alpha}_1^{\mathrm{wls}} - \widehat{\alpha}_0^{\mathrm{wls}} + \bar{X}^\top\left(\widehat{\gamma}_1^{\mathrm{wls}} - \widehat{\gamma}_0^{\mathrm{wls}}\right)$$

What is the interpretation?

$$\mu_i(x_i, \hat{\alpha}_i^{unc}, \hat{\gamma}_i^{unc}) = \hat{\alpha}_i^{unc} + \hat{\gamma}^{unc} x$$

$$w_i = \begin{cases} \frac{1}{e(x_i)} & z_i = 1 \\ \frac{1}{1 - e(x_i)} & z_i = 0 \end{cases}$$

$z \sim Z$ with $w_i \longrightarrow IPW$

$y \sim x + z + xz$ with $w_i \longrightarrow DR$

$\mu_1(x)$

Weighted Regression with Propensity

1. Reweight $Y \sim Z$ with propensity leads to valid estimator; i.e. IPW;
2. Reweight $Y \sim Z + X + ZX$ with propensity leads to doubly robust estimator; i.e. the outcome model can be misspecified.

$\mu_i(x)$. β_2^{new} β_2 $\longrightarrow$ β_2

$e(x)$ $\dfrac{E^{z_i(y_i - \alpha - x_i\beta)}}{e(x_i)}$

$= E\ y_i(1) - x_i\beta$

$\cdots$ $E\ y_i(0) - x_i\beta$

Structural model

- Model $Y_i(z)$ instead of Y_i
  - Estimation involves $Y_i(z)$, e.g., MLE, moment estimation
- Under ignorability, for any function h

$$\mathbb{E}\left\{h\left(Y_i(1), X_i\right)\right\} = \mathbb{E}\left\{\frac{Z_i h\left(Y_i, X_i\right)}{e\left(X_i\right)}\right\}$$

$$\mathbb{E}\left\{h\left(Y_i(0), X_i\right)\right\} = \mathbb{E}\left\{\frac{\left(1 - Z_i\right) h\left(Y_i, X_i\right)}{1 - e\left(X_i\right)}\right\}$$

- IPW corresponds to $h(y, x) = y$
- Estimation involves only Y_i with inverse propensity score weighting

$$h(y,x) = (y-a)^2$$

$$h(y,x) = (y-a-xr)^2$$

Regression with propensity score as a covariate

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

1. $E(Y_i(1) - Y_i(0)|e(X_i)) = E(Y_i|e(X_i), Z_i = 1) - E(Y_i|e(X_i), Z_i = 0)$
2. Regress $Y_i \sim Z_i + e(X_i) + Z_i e(X_i)$
3. Can we do more?

$$E(Y \mid X, Z=1) = \mu_1(X)$$

$$\mu_1^0(X) = \lambda r$$

$$Y \sim \mu_1^0(X) + \frac{r}{e(X)}$$

$$Y \sim X + \varepsilon + \lambda Z$$

$$Y \sim X + Z + XZ + \frac{1}{e(X)} + \frac{Z}{e(X)}$$

Regression with propensity score as a covariate

- Under ignorability, $Z_i \perp \{Y_i(1), Y_i(0)\} \mid e(X_i)$
  - include propensity score (spline basis) in the outcome model (Little and An, 2004; Zhang and Little, 2009)
  - include $1/e(X)$ as regressor in the outcome model (Scharfstein, 1997; Bang and Robins, 2005) – why and is it different from $e(X)$?
- Procedure for estimating $\mathbb{E}\{Y_i(1)\}$:
  - fit a model for propensity score ~~logit $\{e(X_i)\} = X_i^\top \beta$~~, obtain $\hat{e}(X_i)$
  - fit ~~a linear regression~~ model of $Y_i(1)$ on X_i and $\hat{e}(X_i)$ ~~$\mathbb{E}(Y_i(1) \mid X_i) = X_i^\top \gamma + \phi/\hat{e}(X_i)$~~, obtain ~~$\hat{\gamma}$ and $\hat{\phi}$~~
  - calculate the estimator for $\mathbb{E}\{Y_i(1)\}$

$$\hat{\mu}_R = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}\{Y(1) \mid X_i\} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^\top \hat{\gamma} + \frac{\hat{\phi}}{\hat{e}(X_i)}\right)$$

$$Y \sim \sum f(X_i) + \frac{1}{e(X_i)}$$

$$\downarrow$$

$$E(Y_{i(1)} | X)$$

Double robustness

- From the property of linear regression

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{\hat{e}\left(X_i\right)}\left\{Y_i-\left(X_i^{\top}\hat{\gamma}+\frac{\hat{\phi}}{\hat{e}\left(X_i\right)}\right)\right\}=0 \tag{1}$$

- $\hat{\mu}_R$ has the same expression as the doubly robust estimator

$$\hat{\mu}_R=\frac{1}{n}\sum_{i=1}^{n}\left(X_i^{\top}\hat{\gamma}+\frac{\hat{\phi}}{\hat{e}\left(X_i\right)}\right)+\sum_{i=1}^{n}\frac{Z_i}{\hat{e}\left(X_i\right)}\left\{Y_i-\left(X_i^{\top}\hat{\gamma}+\frac{\hat{\phi}}{\hat{e}\left(X_i\right)}\right)\right\}$$

- $\hat{\mu}_R$ is consistent ~~if either~~ outcome or propensity score model is correctly specified

$$\hat{\mu}_{1,DR} = \frac{1}{n}\sum \frac{z_i(y_i - x_i\hat{\beta} - \frac{\phi}{e(x_i)})}{\hat{e}(x_i)}$$

$$+ \frac{1}{n}\sum\left(x_i\hat{\beta} + \frac{\hat{\phi}}{e(x_i)}\right)$$

$$= \hat{\mu}_R$$

Generalization

- Regression estimator becomes doubly robust estimator under Equation (1)
  - linear model with $1/\hat{e}(X_i)$ as regressor guarantees Equation (1) – how about with $\hat{e}(X_i)$
  - $X_i^\top \gamma$ determines the outcome model
- Generalization: $\mathbb{E}\{Y_i(1) \mid X_i\} = \hat{\mathbb{E}}^0\{Y_i(1) \mid X_i\} + \phi/\hat{e}(X_i)$, where $\hat{\mathbb{E}}^0\{Y_i(1) \mid X_i\}$ is an outcome regression estimator – what's the benefit?
- $\hat{\mu}_R$ is consistent if either $\hat{\mathbb{E}}^0\{Y_i(1) \mid X_i\}$ or $\hat{e}(X_i)$ is consistent

Targeted maximum likelihood estimation (TMLE)

- Estimate $\mathbb{E}\{Y_i(1) \mid X_i\}$ to get initial estimator $\hat{\mathbb{E}}^0\{Y_i(1) \mid X_i\}$
- Estimate propensity score to get $\hat{e}(X_i)$
- Construct a "clever" covariate: $1/\hat{e}(X_i)$
- Update the outcome model to $\hat{\mathbb{E}}^1(Y(1) \mid X)$ by fitting $\mathbb{E}\{Y_i(1) \mid X_i\} = \hat{\mathbb{E}}^0\{Y_i(1) \mid X_i\} + \phi/\hat{e}(X_i)$
- Estimator: $\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}^1\{Y_i(1) \mid X_i\}$
- Van der laan and Rubin (2006) propose TMLE
  - similar procedure for ACE – where is the difference?
  - link function can be used for different types of outcome (e.g., logit link for binary outcome)

Calibration methods

We do not have to model propensity score. Instead, we can figure out a pseudo one.

- Balancing property of $e(X_i)$
- Directly estimate the weights w_i without imposing model
  - weights w_i should satisfy balancing property

$$\mathbb{E}\{w_i Z_i h(X_i)\} = \mathbb{E}\{w_i (1 - Z_i) h(X_i)\} = \mathbb{E}\{h(X_i)\}$$

  - ACE estimation

$$\text{ACE} = \mathbb{E}\{w_i Z_i Y_i\} - \mathbb{E}\{w_i (1 - Z_i) h(X_i)\}$$

- Weights are not unique (e.g. propensity score) $\rightsquigarrow$ minimize some functions of the weights
  - e.g., weights to minimize the variance of the ACE estimator
  - implementation using optimization

$$\sum \frac{z_i h(x_i)}{e(x_i)} = \sum \frac{(1-z_i) h(x_i)}{1-e(x_i)}$$

$w_i = e(x_i)$

find w_i: $\sum \frac{z_i h(x_i)}{w_i} \approx \sum \frac{(1-z_i) h(x_i)}{1-w_i}$

minimize variance or other criteria

$$\frac{1}{n} \sum \frac{z_i \xi_i}{w_i} - \frac{1}{n} \sum \frac{(1-w_i) \xi_i}{1-w_i}$$

Calibration Methods

- Entropy balancing (Hainmueller. 2012. Political Anal.)

$$\left\{w_1^*, w_2^*, \ldots, w_{n_0}^*\right\} = \underset{w}{\operatorname{argmin}} \sum_{i:Z_i=0} w_i \log\left(w_i/q_i\right)$$

subject to

$$w_i \geqslant 0, \quad \sum_{i:Z_i=0} w_i = 1, \quad \sum_{i:Z_i=0} w_i f\left(\mathbf{X}_i\right) = \frac{1}{n}\sum_i f\left(\mathbf{X}_i\right)$$

- Stable weights (Zubizarreta, 2015, JASA)

$$\{w_1^*, w_2^*, \ldots, w_n^*\} = \underset{w}{\operatorname{argmin}} \|\mathbf{w} - \overline{\mathbf{w}}\|_2^2$$

subject to

$$w_i \geqslant 0, \quad \sum_{i:Z_i=0} w_i = 1, \quad \left|\sum_{i:Z_i=0} w_i X_{ij} - \frac{1}{n}\sum_i X_{ij}\right| \leqslant \delta_j$$

Weighted OLS with PS

① weight $\sim Z$ with PS $\rightarrow$ IPW

② weight $\sim X + Z + XZ$ with PS

$\rightarrow$ doubly robust estimator

Structural model

Add PS into OR

$$Y \sim X + \frac{1}{e(X)} \mid Z=1$$

$$Y \sim \hat{E}^0(Y(1)|X) + \frac{1}{e(X)} \mid Z=1$$

model-free method

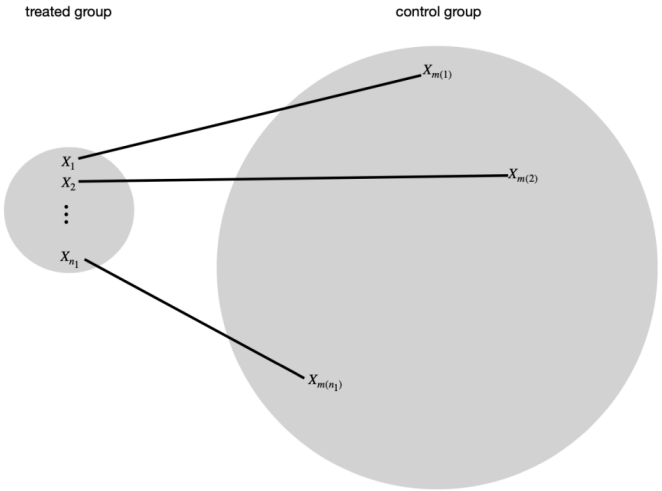
Calibration

Matching

What if we do not want to model the outcome or propensity score

- Popular in empirical research
- Easy implementation without modeling
- Complicated theory – why?
- A practical guide for matching: Stuart (2010) "Matching methods for causal inference: A review and a look forward"

Exact matching



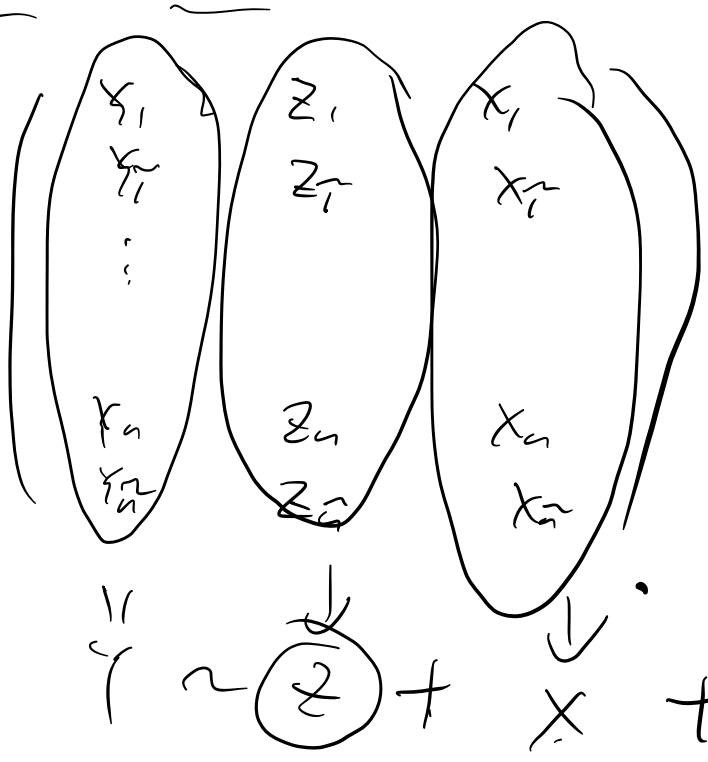
- Exact matching $\rightsquigarrow$ perfect covariate balance
- No model dependence
- Infeasible when there are many covariates and covariates are continuous

Matching based on distance measures

- Matching based on distances between covariates $d(\mathbf{X}_i, \mathbf{X}_j)$
  - Euclidean distance $d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_2^2$
  - Mahalanobis distance $d(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i - \mathbf{X}_j)^\top \Omega^{-1} (\mathbf{X}_i - \mathbf{X}_j)$
  - Propensity score
- Different types of matching
  - One to one, one to many
  - With or without replacement (non i.i.d vs i.i.d)
  - Optimal matching
  - Full matching
- For a given unit in the treatment (control) group, find the unit with the smallest distance in the control (treatment) group
  - 1 to M matching ($M > 1$)
  - depends on the distance measure
  - may drop units that are hard to find matches
  - adjustment after matching – regression or propensity?

Connection with Regression

$(Y_i, \tilde{Y}_i)$



Covariate adjustment in RCT

$$\frac{1}{n}\sum \left\{ Z_i (Y_i - \tilde{Y}_i) + (1 - Z_i)(\tilde{Y}_i - Y_i) \right\}$$

Matching estimator for the ACE

- For a treated unit i, we find the M matched units in the control group and denote $\tilde{Y}_i = \frac{1}{M}\sum_{k\in\mathcal{M}_i} Y_k$, where $\mathcal{M}_i$ is the set of matched units from the control group for unit i
- For a control unit i, we find the M matched units in the treatment group and denote $\tilde{Y}_i = \frac{1}{M}\sum_{k\in\mathcal{M}_i} Y_k$, where $\mathcal{M}_i$ is the set of matched units from the treatment group for unit i
- Matching estimator

$$\widehat{\text{ACE}}^{\text{matching}} = \frac{1}{n}\left\{\sum_{i:Z_i=1}\left(Y_i - \tilde{Y}_i\right) + \sum_{i:Z_i=0}\left(\tilde{Y}_i - Y_i\right)\right\}$$

Matching estimator for the ACE

- Abadie and Imbens (2006, 2008, 2011) study the properties of the matching estimator with replacement assuming $(Z_i, X_i, Y_i(1), Y_i(0))$ are i.i.d.
- $\widehat{\text{ACE}}^{\text{matching}}$ has non-negligible bias especially when X is multidimensional – why?
- Estimator for the bias $\hat{B} = n^{-1} \sum_{i=1}^{n} \hat{B}_i$

$$\hat{B}_i = \begin{cases} M^{-1} \sum_{k \in \mathcal{M}_i} \{\hat{\mu}_0(X_i) - \hat{\mu}_0(X_k)\} & Z_i = 1 \\ M^{-1} \sum_{k \in \mathcal{M}_i} \{\hat{\mu}_1(X_k) - \hat{\mu}_1(X_i)\} & Z_i = 0 \end{cases}$$

  - $\{\hat{\mu}_1(X_i), \hat{\mu}_0(X_i)\}$ are the predicted outcomes, e.g., from OLS
  - bias corrected estimator: $\widehat{\text{ACE}}^{\text{mbc}} = \widehat{\text{ACE}}^{\text{matching}} - \hat{B}$

$$\gamma_i = 1 \quad \frac{1}{M} \sum_{k \in M_i} \left\{ \hat{\mu}_0(X_i) - \hat{\mu}_0(X_k) \right\}$$

$$= \hat{\mu}_0(X_i) - \frac{1}{M} \sum_{k \in M_i} \hat{\mu}_0(X_k)$$

$$\left(\gamma_i - \frac{1}{M} \sum_{k \in M_i} \gamma_k \right) - \beta_i$$

$$= \gamma_i(1) - \frac{1}{M} \sum_{k \in M_i} \gamma_k(0) - \beta_i$$

$$= \gamma_i(1) - \hat{\mu}_0(X_i) - \frac{1}{M} \sum \left(\gamma_k(0) - \hat{\mu}_0(X_k) \right)$$

$$\gamma_i(1) - \hat{\mu}_0(X_i)$$

$$\gamma_k(0) - \hat{\mu}_0(X_k) = \hat{\varepsilon}_{k,0}$$

$$\gamma_i(1) - \mu_0(X_i)$$

Matching estimator for the ACE

- Linear expansion form: $\widehat{\text{ACE}}^{\text{mbc}} = n^{-1} \sum_{i=1}^{n} \hat{\psi}_i$

$$\hat{\psi}_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (2Z_i - 1)(1 + K_i/M)\{Y_i - \hat{\mu}_{Z_i}(X_i)\}$$

  - K_i is the number of times that unit i is used in a match

- Variance estimator

$$\widehat{\text{var}}\left\{\widehat{\text{ACE}}^{\text{mbc}}\right\} = \frac{1}{n^2} \sum_{i=1}^{n} \left(\hat{\psi}_i - \widehat{\text{ACE}}^{\text{mbc}}\right)^2$$

- Otsu and Rai (2017) propose a bootstrap procedure based on the linear expansion

Connection with doubly robust estimator

- $\widehat{\text{ACE}}^{\text{mbc}}$ is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_1\left(X_i\right)-\hat{\mu}_0\left(X_i\right)\right\}+\frac{1}{n}\sum_{i=1}^{n}\left\{\left(1+\frac{K_i}{M}\right)Z_i\hat{R}_i-\left(1+\frac{K_i}{M}\right)\left(1-Z_i\right)\hat{R}_i\right\}$$

  - $\hat{R}_i = Y_i - \hat{\mu}_{Z_i}\left(X_i\right)$ is the residual from outcome regression
  - $n^{-1}\sum_{i=1}^{n}\left\{\hat{\mu}_1\left(X_i\right)-\hat{\mu}_0\left(X_i\right)\right\}$ is the outcome regression estimator
  - the second term is similar to the inverse probability weighting of residual R_i
- Matching can be viewed as a nonparametric method to estimate the propensity score
  - $1 + K_i/M$ should be similar to $1/\hat{e}\left(X_i\right)$ for $Z_{=}1$ and $1/\left\{1-\hat{e}\left(X_i\right)\right\}$ for $Z_i = 0$ – why?
  - Lin et al. (2023) provide a formal theory

$$\frac{1}{n} \sum \left\{ \frac{Z_i (Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1 - Z_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right\}$$

$$\frac{M_i}{M + k_i} \sim \hat{e}(X_i) \quad \text{for } i \text{ such that } Z = 1$$

Summary

- The identification of the average treatment effect in observational studies typically requires:
  - overlap
  - ignorability
- Various methods to estimate causal effects:
  - use propensity score in regression
    - connection to DR estimator
  - Calibration methods
    - estimate weights without modeling
    - need to choose which covariates to balance
  - Matching
    - easy implementation but complicated theory
    - non-negligible bias $\rightsquigarrow$ bias corrected matching estimator
    - connection to weighting methods

Suggested readings

- Using propensity score in regression
  - Ding. Chapter 14
  - Robins, 2007. "Comment: Performance of Double-Robust Estimators When "Inverse Probability" Weights Are Highly Variable"
- Calibration methods
  - papers by Jose Zubizarreta and references therein
- Matching
  - DING. Chapter 15
  - papers by Abadie and Imbens
  - Stuart, 2010. "Matching methods for causal inference: A review and a look forward"