CISER Causal Inference Workshop

Session 2: Advanced Confounder Control

Chencheng Cai

Department of Mathematics and Statistics

Washington State University

Mar 26, 2025

Confounder

A **confounder** is a variable that is associated with both the treatment and the outcome.

If we do not control for the confounder, the estimated treatment effect may be biased, because the confounder is unbalanced between the treatment and control groups.

Simpson's Paradox

The batting averages of two baseball players in 1995 and 1996:

Batter \ Year	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

* batting average = hits / at-bats.

Simpson's Paradox

The batting averages of two baseball players in 1995 and 1996:

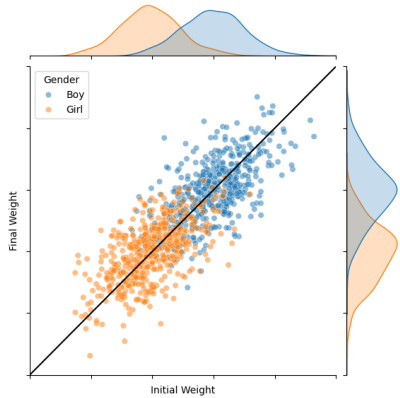
Batter \ Year	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

* batting average = hits / at-bats.

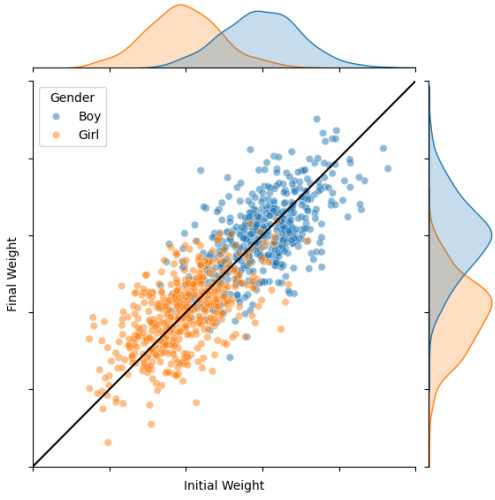
- Each year, player Jeter had a higher batting average than player Justice.
- The overall average of Justice is higher than Jeter.
- Time is a confounder (for batting average and for number of at-bats).

Lord's Paradox

- A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects.
- Two groups: boys vs girls.
- Outcome: weight gain.



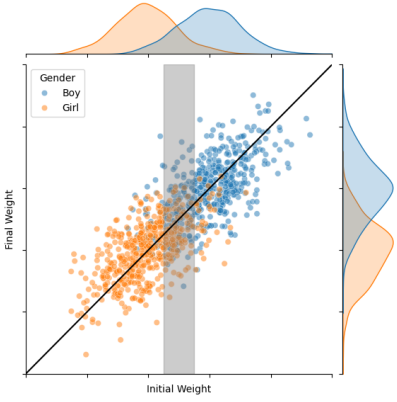
Lord's Paradox

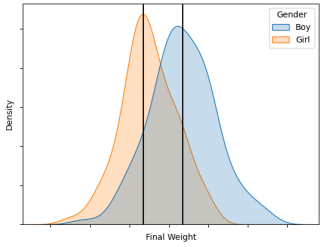


Statistician 1:

- Average weight gain for boys ~ 0.
- Average weight gain for girls ~ 0.
- Conclusion: no significant difference in weight gain between boys and girls.

Lord's Paradox





Statistician 2:

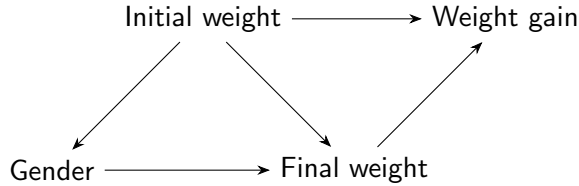
- Choose students with similar initial weights.
- The average weight gains are significantly different for boys and girls.
- Conclusion:
  boys have higher weight gains than girls.

Explanations

- **Holland and Rubin (1983)**:
  - Statistician 1 looks at the average descriptive statistics of the two groups.
  - The two groups in nature have different distributions of initial weights (a confounder).
  - Statistician 2 conditions on the confounder and compares the weight gains.
  - Statistician 2's solution is aligned with the potential outcome framework.

Explanations

- **Holland and Rubin (1983)**:
  - Statistician 1 looks at the average descriptive statistics of the two groups.
  - The two groups in nature have different distributions of initial weights (a confounder).
  - Statistician 2 conditions on the confounder and compares the weight gains.
  - Statistician 2's solution is aligned with the potential outcome framework.
- **Pearl (2014)**:
  Should consider the graphical causal model and do calculus.



Super-Population Perspective and Unconfoundedness

A **super-population perspective** assumes that the observed data are a random sample from a super-population. So are the covariates.

Super-Population Perspective and Unconfoundedness

A **super-population perspective** assumes that the observed data are a random sample from a super-population. So are the covariates.

Unconfoundedness (under super-population perspective) is an assumption on the joint distribution of $(Y_i(1), Y_i(0), W_i, X_i)$:

$$(Y_i(0) \mid X_i, W_i = 1) \stackrel{D}{=} (Y_i(0) \mid X_i, W_i = 0) \quad \text{for all } i$$

and

$$(Y_i(1) \mid X_i, W_i = 1) \stackrel{D}{=} (Y_i(1) \mid X_i, W_i = 0) \quad \text{for all } i.$$

Super-Population Perspective and Unconfoundedness

A **super-population perspective** assumes that the observed data are a random sample from a super-population. So are the covariates.

Unconfoundedness (under super-population perspective) is an assumption on the joint distribution of $(Y_i(1), Y_i(0), W_i, X_i)$:

$$(Y_i(0) \mid X_i, W_i = 1) \stackrel{D}{=} (Y_i(0) \mid X_i, W_i = 0) \quad \text{for all } i$$

and

$$(Y_i(1) \mid X_i, W_i = 1) \stackrel{D}{=} (Y_i(1) \mid X_i, W_i = 0) \quad \text{for all } i.$$

Or, we can write it as

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i.$$

Unconfoundedness

- Unconfoundedness implies the connection between the observed potential outcomes and the missing potential outcomes:

$$\left(Y_i^{mis} \mid X_i, W_i = w\right) \stackrel{D}{=} \left(Y_i^{obs} \mid X_i, W_i = 1 - w\right) \quad \text{for all } i.$$

Unconfoundedness

- Unconfoundedness implies the connection between the observed potential outcomes and the missing potential outcomes:

$$\left(Y_i^{mis} \mid X_i, W_i = w\right) \stackrel{D}{=} \left(Y_i^{obs} \mid X_i, W_i = 1 - w\right) \quad \text{for all } i.$$

- The unconfoundedness assumption is **not** testable.
- The common practice is to include all the pre-treatment covariates in X_i.

Balancing Score

It is ususally impossible to find many pair of units from each of the treated and control groups that have the **same** values of X_i.

Balancing Score

It is ususally impossible to find many pair of units from each of the treated and control groups that have the **same** values of X_i.

Instead, we can find a function of X_i, denoted by $b(X_i)$, called the **balancing score**, such that

$$W_i \perp\!\!\!\perp X_i \mid b(X_i)$$

Balancing Score

It is ususally impossible to find many pair of units from each of the treated and control groups that have the **same** values of X_i.

Instead, we can find a function of X_i, denoted by $b(X_i)$, called the **balancing score**, such that

$$W_i \perp\!\!\!\perp X_i \mid b(X_i)$$

Choices of $b(X_i)$ include:

- X_i itself.
- Propensity score: $e(X_i)$.
- Expert-selected subsets/transformations of X_i.

Balancing Score

Unconfoundedness Given a Balancing Score:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid b(X_i).$$

Balancing Score

Unconfoundedness Given a Balancing Score:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid b(X_i).$$

Coarseness of Balancing Scores:
The propensity score is the coarsest balancing score. That is the propensity score is a function of any other balancing score.

Causal Methods using Balancing Scores

- **Model-based Imputation**: Impute the missing potential outcomes using the covariates.
- **Matching**: Match treated and control units with similar values of the balancing score.
- **Stratification**: Stratify the sample based on the values of the balancing score.
- **Weighting**: Weight the treated and control units.

Regression-based Method

Suppose we consider a linear regression model for the potential outcomes:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} = \begin{pmatrix} X_i\beta_0 \\ X_i\beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix}, \quad \text{with } \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0\sigma_1 \\ \sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right)$$

Regression-based Method

Suppose we consider a linear regression model for the potential outcomes:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} = \begin{pmatrix} X_i\beta_0 \\ X_i\beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix}, \quad \text{with } \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0\sigma_1 \\ \sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right)$$

The coefficients can be estimated by OLS:

$$\hat{\beta}_0 = \arg\min_{\beta} \sum_{i:W_i=0} (Y_i^{obs} - X_i\beta)^2, \quad \hat{\beta}_1 = \arg\min_{\beta} \sum_{i:W_i=1} (Y_i^{obs} - X_i\beta)^2$$

Regression-based Method

Suppose we consider a linear regression model for the potential outcomes:

$$\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} = \begin{pmatrix} X_i\beta_0 \\ X_i\beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix}, \quad \text{with } \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0\sigma_1 \\ \sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right)$$

The coefficients can be estimated by OLS:

$$\hat{\beta}_0 = \arg\min_{\beta} \sum_{i:W_i=0} (Y_i^{obs} - X_i\beta)^2, \quad \hat{\beta}_1 = \arg\min_{\beta} \sum_{i:W_i=1} (Y_i^{obs} - X_i\beta)^2$$

The average treatment effect can be estimated by

$$\hat{\tau}^{ols} = \frac{1}{N} \sum_{i=1}^{N} \left[W_i(Y_i^{obs} - X_i\hat{\beta}_0) + (1 - W_i)(X_i\hat{\beta}_1 - Y_i^{obs}) \right]$$

Regression-based Method

$\hat{\beta}_0$ is **fitted** using $\{X_i : W_i = 0\}$, but $\hat{\beta}_0$ is used in **prediction** for $\{X_i : W_i = 1\}$.

Regression-based Method

$\hat{\beta}_0$ is **fitted** using $\{X_i : W_i = 0\}$, but $\hat{\beta}_0$ is used in **prediction** for $\{X_i : W_i = 1\}$.

- If X_i differs a lot in domains for the two groups, the prediction is extrapolation.

Regression-based Method

$\hat{\beta}_0$ is **fitted** using $\{X_i : W_i = 0\}$, but $\hat{\beta}_0$ is used in **prediction** for $\{X_i : W_i = 1\}$.

- If X_i differs a lot in domains for the two groups, the prediction is extrapolation.
- For completely randomized experiments, the expected difference in X_i is 0 —- not a bit issue of extrapolation.
- For other assigning mechanisms, especially for observational studies, extrapolation is a big issue —- the performances depends strongly on the model specification.

Regression-based Method

$\hat{\beta}_0$ is **fitted** using $\{X_i : W_i = 0\}$, but $\hat{\beta}_0$ is used in **prediction** for $\{X_i : W_i = 1\}$.

- If X_i differs a lot in domains for the two groups, the prediction is extrapolation.
- For completely randomized experiments, the expected difference in X_i is 0 —- not a bit issue of extrapolation.
- For other assigning mechanisms, especially for observational studies, extrapolation is a big issue —- the performances depends strongly on the model specification.
- Solution: the observations should be **weighted** in OLS such that the covariates are **balanced** in distribution between the two groups.

Weighting using Propesnity Score

The inverse propensity score weighting (IPW) is a popular method to balance the covariates.

Weighting using Propesnity Score

The inverse propensity score weighting (IPW) is a popular method to balance the covariates.

Two properties of the propensity score in weighting:

- **Unbiasedness**:

$$\mathbb{E}\left(\frac{W_i Y_i^{obs}}{e(X_i)}\right) = \mathbb{E}(Y_i(1)), \quad \mathbb{E}\left(\frac{(1-W_i)Y_i^{obs}}{1-e(X_i)}\right) = \mathbb{E}(Y_i(0)),$$

  where the expectation is taken over the super-population distribution.

Weighting using Propesnity Score

The inverse propensity score weighting (IPW) is a popular method to balance the covariates.

Two properties of the propensity score in weighting:

- **Unbiasedness**:

$$\mathbb{E}\left(\frac{W_i Y_i^{obs}}{e(X_i)}\right) = \mathbb{E}(Y_i(1)), \quad \mathbb{E}\left(\frac{(1-W_i)Y_i^{obs}}{1-e(X_i)}\right) = \mathbb{E}(Y_i(0)),$$

  where the expectation is taken over the super-population distribution.

- **Balanced Covariates**:

$$\frac{p(X_i \mid W_i = 1)}{p(X_i \mid W_i = 0)} \propto \frac{e(X_i)}{1-e(X_i)} \implies \frac{p(X_i \mid W_i = 1)}{e(X_i)} \propto \frac{p(X_i \mid W_i = 0)}{1-e(X_i)}$$

  The density of IPW-ed X_i is the same for the two groups.

Weighting using Propesnity Score

The IPW estimator of the average treatment effect is

$$\hat{\tau}^{ipw} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i Y_i^{obs}}{e(X_i)} - \frac{(1-W_i)Y_i^{obs}}{1-e(X_i)}\right)$$

Weighting using Propesnity Score

The IPW estimator of the average treatment effect is

$$\hat{\tau}^{ipw} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i Y_i^{obs}}{e(X_i)} - \frac{(1-W_i)Y_i^{obs}}{1-e(X_i)}\right)$$

- Advantage: always unbiased under unconfoundedness.
- Disadvantage: the variance of the estimator can be large — due to extreme propensity scores.
- It is also called the **Horvitz-Thompson estimator** in survey sampling.
- **Caution**: $e(X_i)$ still remains to be estimated.

Weighting using Propesnity Score

A few ways to solve extreme propensity scores.

- **Trimming**: exclude the units with extreme propensity scores. Drawback: lose information.
- **Truncation**: saturate the propensity scores at a certain level. Drawback: added bias.
- **Hájek Estimator**: reduces variance by normalizing weights.

$$\hat{\tau}^{hajek} = \frac{\sum_i W_i Y_i^{obs}/e(X_i)}{\sum_i W_i/e(X_i)} - \frac{\sum_i (1-W_i) Y_i^{obs}/(1-e(X_i))}{\sum_i (1-W_i)/(1-e(X_i))}$$

Drawback: added bias for finite sample.

Subclassification on Propensity Scores

We divide the domain of the propensity score into J strata:

$$b_0 = 0 < b_1 < b_2 < \cdots < b_{J-1} < b_J = 1.$$

Subclassification on Propensity Scores

We divide the domain of the propensity score into J strata:

$$b_0 = 0 < b_1 < b_2 < \cdots < b_{J-1} < b_J = 1.$$

For each stratum j, the difference-in-means is

$$\hat{\tau}^{dif}(j) = \frac{1}{N_t(j)} \sum_{i:b_{j-1}<e(X_i)<b_j} W_i Y_i^{obs} - \frac{1}{N_c(j)} \sum_{i:b_{j-1}<e(X_i)<b_j} (1 - W_i) Y_i^{obs}$$

The overall estimator is

$$\hat{\tau}^{strat} = \sum_{j=1}^{J} \frac{N(j)}{N} \hat{\tau}^{dif}(j)$$

Subclassification on Propensity Scores

We divide the domain of the propensity score into J strata:

$$b_0 = 0 < b_1 < b_2 < \cdots < b_{J-1} < b_J = 1.$$

For each stratum j, the difference-in-means is

$$\hat{\tau}^{dif}(j) = \frac{1}{N_t(j)} \sum_{i:b_{j-1}<e(X_i)<b_j} W_i Y_i^{obs} - \frac{1}{N_c(j)} \sum_{i:b_{j-1}<e(X_i)<b_j} (1 - W_i) Y_i^{obs}$$

The overall estimator is

$$\hat{\tau}^{strat} = \sum_{j=1}^{J} \frac{N(j)}{N} \hat{\tau}^{dif}(j)$$

- Advantage: the variance is reduced by stratification.
- Disadvantage: potential bias due to the discretization of the propensity score.

Subclassification on Propensity Scores

For each stratum j, the bias comes from

- We use $\bar{Y}_t^{obs}(j)$ to estimate $\mathbb{E}(Y_i(1) \mid W_i = 0)$.
- We use $\bar{Y}_c^{obs}(j)$ to estimate $\mathbb{E}(Y_i(0) \mid W_i = 1)$.

They are not equal because the covariates are not perfectly balanced in each stratum.

Subclassification on Propensity Scores

For each stratum j, the bias comes from

- We use $\bar{Y}_t^{obs}(j)$ to estimate $\mathbb{E}(Y_i(1) \mid W_i = 0)$.
- We use $\bar{Y}_c^{obs}(j)$ to estimate $\mathbb{E}(Y_i(0) \mid W_i = 1)$.

They are not equal because the covariates are not perfectly balanced in each stratum.

Bias Correction:

- We fit a parametric model for the potential outcomes:

$$Y_i(W_i) = f(X_i, W_i; \theta) + \epsilon_i.$$

  The model can be as simple as linear.
- We estimate the difference between $\mathbb{E}(Y_i(1) \mid W_i = 0)$ and $\mathbb{E}(Y_i(1) \mid W_i = 0)$ using the fitted model and the discrepancy in the covariates.
- The bias is substrated from the stratified estimator.

Matching Method

Consider an experiment with more treated than control units. The **matching** method is to find for each treated unit a **similar** control unit.

Matching Method

Consider an experiment with more treated than control units. The **matching** method is to find for each treated unit a **similar** control unit.

- **closeness**:
  - **Exact Matching**: the treated and control units have the same values of the covariates. Often for categorical variables.
  - **Metric Matching**: the treated and control units are close in the covariate space. Often for continuous variables. E.g. using Mahalanobis distance.
  - **Propensity Score Matching**: the treated and control units have similar propensity scores.

$$d(i,j) = \left[\log \frac{e(X_i)}{1-e(X_i)} - \log \frac{e(X_j)}{1-e(X_j)}\right]^2$$

Matching Method

Consider an experiment with more treated than control units. The **matching** method is to find for each treated unit a **similar** control unit.

- **closeness**:
  - **Exact Matching**: the treated and control units have the same values of the covariates. Often for categorical variables.
  - **Metric Matching**: the treated and control units are close in the covariate space. Often for continuous variables. E.g. using Mahalanobis distance.
  - **Propensity Score Matching**: the treated and control units have similar propensity scores.

$$d(i,j) = \left[\log \frac{e(X_i)}{1-e(X_i)} - \log \frac{e(X_j)}{1-e(X_j)}\right]^2$$

- Replacement:
  - **With Replacement**: the control unit can be matched to multiple treated units.
  - **Without Replacement**: the control unit can be matched to only one treated unit.

Matching Method

Algorithms:

- **Greedy Matching 1**: for each treated unit, find the available closest control unit. Repeat until all treated units are matched.
- **Greedy Matching 2**: within all avaialble pairs, find the closest pair. Repeat until all treated units are matched.
- **Optimal Matching**: find the best matching pairs that minimize the total distance. (NP-hard, optimal transportations)

Matching Method

Algorithms:

- **Greedy Matching 1**: for each treated unit, find the available closest control unit. Repeat until all treated units are matched.
- **Greedy Matching 2**: within all avaialble pairs, find the closest pair. Repeat until all treated units are matched.
- **Optimal Matching**: find the best matching pairs that minimize the total distance. (NP-hard, optimal transportations)

Common Practices:

- Poor matches are often excluded.
- Could also use one-to-many matching using a caliper.
- Balance of the covariates should be checked after matching.
- **Cautious**: distribution shift after matching!!!

Matching Method

The estimator is

$$\hat{\tau}^{match} = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i^{obs} - Y_{m(i)}^{obs}),$$

where $m(i)$ is the matched control unit for treated unit i.

Matching Method

The estimator is

$$\hat{\tau}^{match} = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i^{obs} - Y_{m(i)}^{obs}),$$

where $m(i)$ is the matched control unit for treated unit i.

Caution: because the matching is per treated unit, the distribution of covariates for treated can be different from that of the population.

Matching Method

The estimator is

$$\hat{\tau}^{match} = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i^{obs} - Y_{m(i)}^{obs}),$$

where $m(i)$ is the matched control unit for treated unit i.

Caution: because the matching is per treated unit, the distribution of covariates for treated can be different from that of the population.

We are estimating the **average treatment effect for the treated units**, not for the population.

- **Finite Sample Version**:

$$\tau_{fs,t} = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i(1) - Y_i(0))$$

- **Super-Population Version**:

$$\tau_{sp,t} = \mathbb{E}(Y_i(1) - Y_i(0) \mid W_i = 1)$$

Matching Method

If the matching is **exact**, the estimator is unbiased.

$$E(Y_i^{obs} - Y_{m(i)}^{obs} \mid X_i) = E(Y_i(1) - Y_i(0) \mid X_i).$$

The variance of the estimator $\hat{\tau}_t^{match}$ can be estimated from the sample variance of the differences.

Matching Method

If the matching is **exact**, the estimator is unbiased.

$$E(Y_i^{obs} - Y_{m(i)}^{obs} \mid X_i) = E(Y_i(1) - Y_i(0) \mid X_i).$$

The variance of the estimator $\hat{\tau}_t^{match}$ can be estimated from the sample variance of the differences.

If the matching is **inexact**, the estimator is biased.

- The bias is due to the difference in the covariates/propensity scores between the treated and control units.
- The bias can be corrected by the regression adjustment.

Matching Method

If the matching is **exact**, the estimator is unbiased.

$$E(Y_i^{obs} - Y_{m(i)}^{obs} \mid X_i) = E(Y_i(1) - Y_i(0) \mid X_i).$$

The variance of the estimator $\hat{\tau}_t^{match}$ can be estimated from the sample variance of the differences.

If the matching is **inexact**, the estimator is biased.

- The bias is due to the difference in the covariates/propensity scores between the treated and control units.
- The bias can be corrected by the regression adjustment.

One-to-many matching can be used to reduce bias and variance of the estimator.

- Bias is reduced because multiple control units are matched to a treated unit — smaller discrepancy in the averaged covariates.
- Variance is reduced because the sample size is increased. But the improvement in variance is limited.

Matching Method

To estimate the average treatment effect for the control $\tau_{sp,c}$, we can match all controlled units instead.

Matching Method

To estimate the average treatment effect for the control $\tau_{sp,c}$, we can match all controlled units instead.

To estimate the average treatment effect for the population τ_{sp}, we need to estimate both $\tau_{sp,t}$ and $\tau_{sp,c}$ and combine them.

$$\tau_{sp} = \frac{N_t}{N}\tau_{sp,t} + \frac{N_c}{N}\tau_{sp,c}$$

Other Technical Issues

- All the propensity scores need to be estimated in the first place. The most common model is a logistic regression model:

$$\log \frac{P(W_i = 1 \mid X_i)}{1 - P(W_i = 1 \mid X_i)} = X_i \beta.$$

- The model specification is crucial. The model should be flexible enough to capture the relationship between the covariates and the treatment assignment.
- Besides the caual effect estimators, the standard errors should also be estimated. We skip the details here.
- The hypothesis testing for the average treatment effect usually follows a Z-test (when the sample size is large).

Example

We take one example from Imbens and Rubins' book to demonstrate the techniques in balancing the covariates.

- Treatment: prenatal exposure to barbiturates
- Covariates: quite a few regarding the born child and the mother.
- Outcome: cognitive development measured many years later
- 745 treated and 7198 control.
- It is an observational study.

Example

Label	Variable Description	Controls (N_c =7198)		Treated (N_t =745)		t-Stat Difference
		Mean	(S.D.)	Mean	(S.D.)	
sex	Sex of child (female is 0)	0.51	(0.50)	0.50	(0.50)	−0.3
antih	Exposure to antihistamine	0.10	(0.30)	0.17	(0.37)	4.5
hormone	Exposure to hormone treatment	0.01	(0.10)	0.03	(0.16)	2.5
chemo	Exposure to chemotherapy agents	0.08	(0.27)	0.11	(0.32)	2.5
cage	Calendar time of birth	−0.00	(1.01)	0.03	(0.97)	0.7
cigar	Mother smoked cigarettes	0.54	(0.50)	0.48	(0.50)	−3.0
lgest	Length of gestation (10 ordered categories)	5.24	(1.16)	5.23	(0.98)	−0.3
lmotage	Log of mother's age	−0.04	(0.99)	0.48	(0.99)	13.8
lpbc415	First pregnancy complication index	0.00	(0.99)	0.05	(1.04)	1.2
lpbc420	Second pregnancy complication index	−0.12	(0.96)	1.17	(0.56)	55.2
motht	Mother's height	3.77	(0.78)	3.79	(0.80)	0.7
motwt	Mother's weight	3.91	(1.20)	4.01	(1.22)	2.0
mbirth	Multiple births	0.03	(0.17)	0.02	(0.14)	−1.9
psydrug	Exposure to psychotherapy drugs	0.07	(0.25)	0.21	(0.41)	9.1
respir	Respiratory illness	0.03	(0.18)	0.04	(0.19)	0.7
ses	Socioeconomic status (10 ordered categories)	−0.03	(0.99)	0.25	(1.05)	7.0
sib	If sibling equal to 1, otherwise 0	0.55	(0.50)	0.52	(0.50)	−1.6

Example

Step 1: Estimate the propensity score.

- The propensity score is estimated by a logistic regression model.
- Due to the large number of parameters and possible interactions, the logistic regression model selects covariates by a sequential manner.

Example

Variable	EST	$\widehat{(\text{s.e.})}$	t-Stat
Intercept	−5.67	(0.23)	−24.4
Linear terms			
sex	0.12	(0.09)	1.3
lmotage	0.52	(0.11)	4.7
ses	0.06	(0.09)	0.6
lpbc420	2.37	(0.36)	6.6
mbirth	−2.11	(0.36)	−5.9
chemo	−3.51	(0.67)	−5.2
psydrug	−3.37	(0.55)	−6.1
sib	−0.24	(0.22)	−1.1
cage	−0.56	(0.26)	−2.2
lgest	0.57	(0.23)	2.5
motwt	0.49	(0.17)	2.9
cigar	−0.15	(0.10)	−1.5
antih	0.17	(0.13)	1.3
Second-order terms			
lpbc420 × sib	0.60	(0.19)	3.1
motwt× motwt	−0.10	(0.02)	−4.5
lpbc420 × psydrug	1.88	(0.39)	4.8
ses× sib	−0.22	(0.10)	−2.2
cage× antih	−0.39	(0.14)	−2.8
lpbc420× chemo	1.97	(0.49)	4.0
lpbc420 × lpbc420	−0.46	(0.14)	−3.3
cage × lgest	0.15	(0.05)	3.0
lmotage × lpbc420	−0.24	(0.10)	−2.5
mbirth× cage	−0.88	(0.39)	−2.3
lgest × lgest	−0.04	(0.02)	−2.0
ses× cigar	0.20	(0.09)	2.2
lpbc420× motwt	0.15	(0.07)	2.0
chemo × psydrug	−0.93	(0.46)	−2.0
lmotage× ses	0.10	(0.05)	1.9
cage × cage	−0.10	(0.05)	−1.8
mbirth × chemo	−∞	(0.00)	−∞

Example

Step 2: Stratification.

- The stratification is based on the propensity score.
- An adaptive procedure is conducted:
  - Starting from one stratum, the balance of the covariates is checked.
  - If the balance is not achieved, the stratum is split into two new stratum of equal size.
  - Repeat until all strata are balanced.

Example

Step	Block	Lower Bound	Upper Bound	Width	# Controls	# Treated	t-Stat
1	1	0.00	0.94	0.94	4462	742	36.3
2	**1**	0.00	0.06	0.06	2540	61	3.2
	2	0.06	0.94	0.88	1922	681	23.7
3	**1**	0.00	0.02	0.01	1280	20	2.2
	2	0.02	0.06	0.05	1260	41	0.5
	3	0.06	0.20	0.14	1163	138	3.9
	4	0.20	0.94	0.74	759	543	10.9
4	**1**	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.94	0.57	301	351	5.6

Example

5	1	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.50	0.13	181	144	2.5
	8	0.50	0.94	0.44	120	207	2.3
6	1	0.00	0.01	0.00	644	6	−0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	−0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.42	0.05	101	61	0.3
	8	0.42	0.50	0.08	80	83	0.7
	9	0.50	0.61	0.11	73	90	0.8
	10	0.61	0.94	0.34	47	117	−0.3

Example

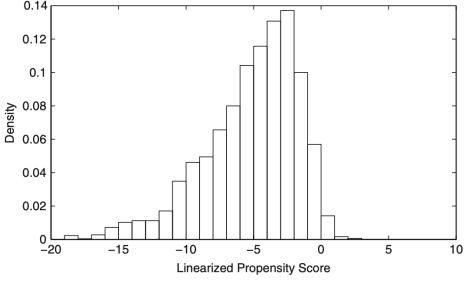
	Within Blocks										Overall		1-Block
	1	2	3	4	5	6	7	8	9	10	t-Test	F-Test (z-Value)	t-Test
Covariate													
sex	−0.05	−2.27	1.97	0.81	0.89	−1.28	0.04	−0.39	−1.42	1.14	0.13	1.22	−0.73
antih	−0.67	−0.47	0.67	0.03	0.37	−0.25	0.38	−0.53	−0.11	0.27	−0.17	−2.88	3.21
hormone	−0.14	−0.42	−0.65	−1.00	0.25	0.71	−0.22	−1.05	−1.10	0.21	−0.99	−0.66	1.66
chemo	0.55	−0.39	−0.78	−0.75	−1.17	1.47	−0.94	0.61	0.66	0.29	−0.27	−0.61	1.76
cage	−1.41	−0.29	−1.04	−0.46	2.11	0.28	0.20	0.46	−1.48	−0.74	−1.38	0.34	1.15
cigar	−0.37	0.55	0.58	1.50	0.31	−0.93	0.21	−0.99	0.25	−0.39	0.52	−1.17	−3.13
lgest	0.90	0.58	−0.07	−0.82	0.79	−0.36	0.05	−0.33	−1.14	1.21	0.71	−1.48	0.12
lmotage	−2.20	−1.37	0.56	1.64	0.95	0.60	−0.96	−1.73	−1.47	0.36	−1.26	1.45	8.56
lpbc415	−0.48	−1.84	−1.00	−0.34	0.59	0.44	−0.20	−0.16	1.07	−0.10	−1.49	−0.82	0.75
lpbc420	1.04	0.84	−0.67	−0.86	−1.61	1.80	−0.39	1.62	1.14	−1.80	0.51	0.59	32.04
motht	−0.84	0.45	−0.67	0.75	0.64	0.09	0.30	−1.37	−0.60	−0.13	−0.50	−1.37	0.90
motwt	1.23	1.14	0.12	−1.23	−0.05	−0.45	−0.32	1.94	−0.01	−0.47	1.08	−0.18	1.44
mbirth	−0.44	−0.80	−1.54	−0.37	1.80	0.20	0.00	2.25	−1.58	−1.60	−1.28	1.00	−2.93
psydrug	−0.66	−1.01	1.05	−0.15	−0.78	0.06	−0.18	0.08	0.09	0.89	−0.29	−1.40	6.32
respir	−0.49	0.53	−0.21	0.98	1.38	0.24	−0.78	−1.51	0.22	−0.28	0.24	−0.49	0.19
ses	−0.60	−0.31	−0.74	1.16	0.82	−0.08	−0.03	−0.82	−0.91	0.36	−0.56	−1.37	5.19
sib	1.42	2.37	−1.09	−1.58	−1.53	0.11	0.63	1.63	1.19	0.23	0.98	1.64	1.48

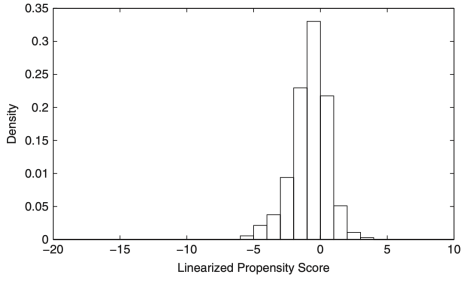
Example

Step 3: Matching Method.

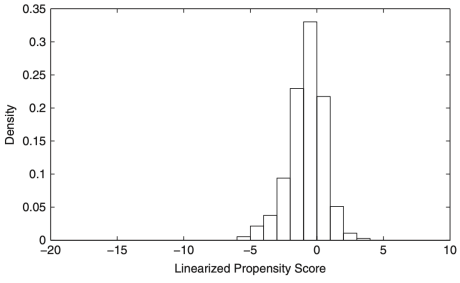
- The matching is based on the Mahalanobis distance or the propensity scores.
- The performance is check by the overlapping scores of the covariates.

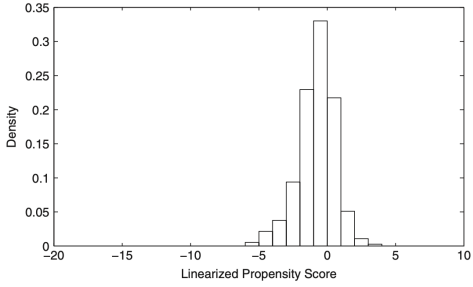
Example


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Density
−20
−15
−10
−5
0
5
10
Linearized Propensity Score


0.35
0.3
0.25
0.2
0.15
0.01
0.05
0
Density
−20
−15
−10
−5
0
5
10
Linearized Propensity Score

Example


Density
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−20
−15
−10
−5
0
5
10
Linearized Propensity Score


Density
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−20
−15
−10
−5
0
5
10
Linearized Propensity Score

Example

	Full Sample				Matched Samples							
					Mahalanobis				Propensity Score			
			$\pi^{0.05}$				$\pi^{0.05}$				$\pi^{0.05}$	
	Nor Dif	Log Rat of STD	Controls	Treated	Nor Dif	Log Rat of STD	Controls	Treated	Nor Dif	Log Rat of STD	Controls	Treated
sex	−0.01	0.00	1.00	1.00	0.00	−0.00	1.00	1.00	−0.03	0.00	1.00	1.00
antih	0.19	0.20	1.00	1.00	0.02	0.01	1.00	1.00	−0.03	−0.02	1.00	1.00
hormone	0.11	0.43	1.00	0.97	0.00	0.00	1.00	1.00	0.01	0.03	1.00	0.97
chemo	0.10	0.14	1.00	1.00	0.00	0.00	1.00	1.00	0.08	0.10	1.00	1.00
cage	0.03	−0.04	0.93	0.97	−0.03	0.03	0.96	0.95	−0.01	−0.00	0.95	0.95
cigar	−0.12	0.00	1.00	1.00	−0.01	−0.00	1.00	1.00	−0.01	−0.00	1.00	1.00
lgest	−0.01	−0.17	0.95	0.98	−0.02	0.13	0.98	0.97	0.00	0.01	0.98	0.97
lmotage	0.53	0.00	0.93	0.93	0.13	0.02	0.97	0.95	0.02	−0.01	0.95	0.97
lpbc415	0.05	0.06	0.99	0.97	0.03	0.06	0.98	0.99	0.07	−0.06	0.99	0.97
lpbc420	1.63	−0.55	0.52	0.72	0.59	−0.01	0.90	0.86	0.10	0.09	0.96	0.94
motht	0.03	0.03	1.00	1.00	−0.03	0.15	1.00	1.00	−0.03	0.03	1.00	1.00
motwt	0.08	0.02	1.00	1.00	0.02	0.09	1.00	1.00	0.05	−0.02	1.00	1.00
mbirth	−0.07	−0.21	0.97	1.00	0.00	0.00	0.98	0.98	0.03	0.12	0.99	0.98
psydrug	0.41	0.47	1.00	1.00	0.00	0.00	1.00	1.00	0.13	0.09	1.00	1.00
respir	0.03	0.07	1.00	1.00	0.00	0.00	1.00	1.00	0.03	0.07	1.00	1.00
ses	0.28	0.06	1.00	1.00	0.03	0.08	0.99	0.96	−0.04	0.02	0.99	0.96
sib	−0.06	0.00	1.00	1.00	0.03	−0.00	1.00	1.00	0.04	−0.00	1.00	1.00
Multivariate measure	0.43				0.24				0.05			

Summary

- **Model Imputation Methods**:
  - Advantages: easy to implement, flexible.
  - Disadvantages: model specification is crucial, extrapolation is a big issue.
- **Weighting Methods**:
  - Advantages: unbiased, balance the covariates.
  - Disadvantages: variance can be large, extreme propensity scores.
- **Stratification Methods**:
  - Advantages: reduce variance, balance the covariates.
  - Disadvantages: bias due to discretization.
- **Matching Methods**:
  - Advantages: balance the covariates, reduce variance.
  - Disadvantages: bias due to inexact matching, distribution shift.

Beyond the Single Causes

The confounder control so far is based on a single treatment (cause). In reality, there could be many (cause, consequence) pairs.

Beyond the Single Causes

The confounder control so far is based on a single treatment (cause). In reality, there could be many (cause, consequence) pairs.

The Book of Why by Judea Pearl
Causal inference in statistics: An overview by Judea Pearl
Causal Inference in Statistics: A Primer by Judea Pearl

Causal Diagram

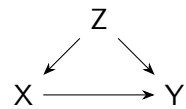
A **causal diagram** is a directed acyclic graph (DAG) that represents the causal relationships between the variables.

- $A \to B$ means A affacts B directly.

Causal Diagram

A **causal diagram** is a directed acyclic graph (DAG) that represents the causal relationships between the variables.

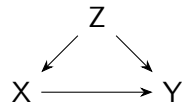
- $A \to B$ means A affacts B directly.



- It is very easy to identify the confounders in the causal diagram.
- For the diagram above, Z is a confounder for the effect of X on Y.

Do Operator

The operator $\mathrm{do}(X = x)$ is used to represent the intervention of setting X to x.



- Conventional probability:

$$P(y, z \mid X = x) = \frac{P(x, y, z)}{P(x)} = \frac{P(y \mid x, z)P(x \mid z)P(z)}{\int P(x \mid z)P(z)dz}$$

- Do operator: (x is no longer random after intervention)

$$P(y, z \mid \mathrm{do}(x)) = P(y \mid x, z)P(z)$$

Do Operator

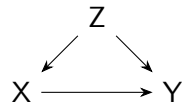
For any DAG causal diagram with independent noise terms, the conditional distribution conditioned on the do operator is

$$P(v_1, \ldots, v_k \mid \mathrm{do}(x_0)) = \prod_{i:v_i \neq X} P(v_i \mid pa_i)|_{x=x_0},$$

where pa_i is the parent set of v_i in the DAG.

Do Operator

Any unmeasured confounders should be integrated out.



We already know

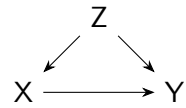
$$P(y, z \mid \mathrm{do}(x)) = P(y \mid x, z)P(z)$$

Therefore,

$$P(y \mid \mathrm{do}(x)) = \sum_{z} P(y \mid x, z)P(z)$$

Do Operator

Any unmeasured confounders should be integrated out.



We already know

$$P(y, z \mid \mathrm{do}(x)) = P(y \mid x, z)P(z)$$

Therefore,

$$P(y \mid \mathrm{do}(x)) = \sum_z P(y \mid x, z)P(z)$$

- As long as we can estimate $P(y \mid x, z)$ and $P(z)$ from the data, we can estimate $P(y \mid \mathrm{do}(x))$.
- The **causal effect** of X on Y is

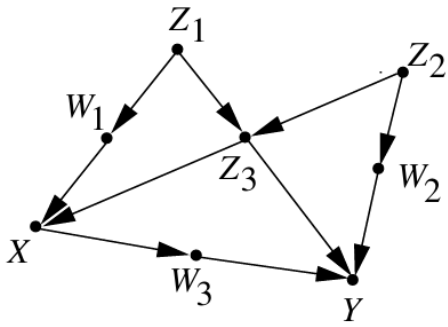
$$\tau = E(Y \mid \mathrm{do}(x = 1)) - E(Y \mid \mathrm{do}(x = 0))$$

Back-door Adjustment

In practice, if we want to estimate the causal effect of X on Y, there could be many confounders. We need to determine all the necessary confounders and adjust for them.

Back-door Adjustment

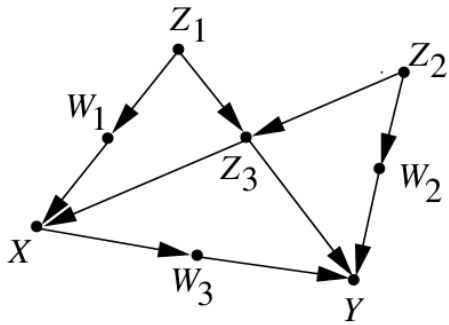
In practice, if we want to estimate the causal effect of X on Y, there could be many confounders. We need to determine all the necessary confounders and adjust for them.



Confounders for the effect of X on Y: Z_1, Z_2, Z_3.

Back-door Adjustment

In practice, if we want to estimate the causal effect of X on Y, there could be many confounders. We need to determine all the necessary confounders and adjust for them.



Confounders for the effect of X on Y: Z_1, Z_2, Z_3.

Such a set of confounders is called a **admissible set**, such that as long as the set is adjusted for, the causal effect is identified.

Back-door Adjustment

Back-door Criterion:
A set S is admissible for adjustment if two conditions hold:

1. No element of S is a descendant of X.
2. S blocks all back-door paths between X and Y.

Back-door Adjustment

Back-door Criterion:
A set S is admissible for adjustment if two conditions hold:

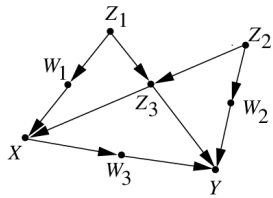
1. No element of S is a descendant of X.
2. S blocks all back-door paths between X and Y.

Back-door Adjustment Formula:
If S is an admissible set, then

$$P(Y = y \mid \mathrm{do}(X = x)) = \sum_{s} P(Y = y \mid X = x, S = s) P(S = s)$$

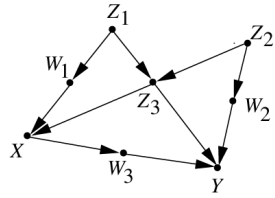
where the summation is over all possible values of S.

Back-door Adjustment



- $\{Z_1, Z_3\}$ is an admissible set.
- $\{Z_3\}$ is an admissible set.
- $\{Z_1, Z_2\}$ is not an admissible set.

Back-door Adjustment



- $\{Z_1, Z_3\}$ is an admissible set.
- $\{Z_3\}$ is an admissible set.
- $\{Z_1, Z_2\}$ is not an admissible set.

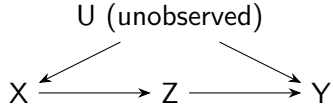
If we adjust for Z_3, the back-door adjustment formula is

$$P(Y = y \mid \mathrm{do}(X = x)) = \sum_{z_3} P(Y = y \mid X = x, Z_3 = z_3) P(Z_3 = z_3)$$

In practice, we only need to estimate the conditional distribution $P(Y \mid X, Z_3)$ and the marginal distribution $P(Z_3)$.

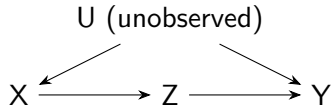
Front-door Adjustment

Suppose we have the following causal diagram. But U is unobserved.



Front-door Adjustment

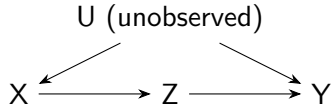
Suppose we have the following causal diagram. But U is unobserved.



- The back-door adjustment does not work because U is unobserved.
- However, the path $X \to U \to Y$ is blocked by U such that U works as a proxy for X.

Front-door Adjustment

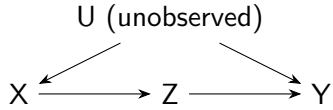
Suppose we have the following causal diagram. But U is unobserved.



- The back-door adjustment does not work because U is unobserved.
- However, the path $X \to U \to Y$ is blocked by U such that U works as a proxy for X.
- The effect of X on Z can be estimated.
- The effect of Z on Y can be estimated.

Front-door Adjustment

Suppose we have the following causal diagram. But U is unobserved.



- The back-door adjustment does not work because U is unobserved.
- However, the path $X \to U \to Y$ is blocked by U such that U works as a proxy for X.
- The effect of X on Z can be estimated.
- The effect of Z on Y can be estimated.
- The effect of X on Y can be estimated by combining the two effects.

Front-door Adjustment

Front-door Criterion:
A set S satisfies the front-door criterion if three conditions hold:

1. S intercepts all directed paths from X to Y.
2. There is no backdoor path from X to S.
3. All backdoor paths from S to Y are blocked by X.

Front-door Adjustment

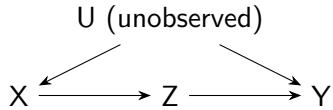
Front-door Criterion:
A set S satisfies the front-door criterion if three conditions hold:

1. S intercepts all directed paths from X to Y.
2. There is no backdoor path from X to S.
3. All backdoor paths from S to Y are blocked by X.

Front-door Adjustment Formula:
If S is an admissible set, then

$$P(Y = y \mid \mathrm{do}(X = x)) = \sum_{s} \sum_{x'} P(Y = y \mid X = x', S = s) P(X = x') P(S = s \mid X = x).$$

Front-door Adjustment



The set $\{Z\}$ satisfies the front-door criterion. The front-door adjustment formula is

$$P(Y = y \mid \mathrm{do}(X = x)) = \sum_{z} \sum_{x'} P(Y = y \mid X = x', Z = z) P(X = x') P(Z = z \mid X = x).$$

Thank you for joining the workshop!

Contact:

Me	chencheng.cai@wsu.edu
CISER	ciser.info@wsu.edu