Matching Methods: An Introduction

NYU Wagner

Rajeev Dehejia

Motivation: smoking and mortality (Cochran, 1968)

Table 1

Death Rates per 1,000 Person-Years

Smoking Group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

*But what other differences among people with **these habits** in **these places** could be affecting **this outcome**?*

Motivation: smoking and mortality (Cochran, 1968)

Table 2

Mean Ages, Years

Smoking Group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

1. Subclassification

To control for differences in age, we would like to compare mortality across different smoking-habit groups **with the same age distribution**

One possibility is to use **subclassification**:

- Divide the smoking-habit samples into age groups
- For each of the smoking-habit samples, calculate the mortality rates for the age groups
- Construct weighted averages of the age groups mortality rates for each smoking-habit sample using a fixed set of weights across all samples (that is, for a fixed distribution of the age variable across all samples)

1. Subclassification: example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total	100	40	40

Question

What would be the average death rate for Pipe Smokers if they had the same age distribution as Non-Smokers?

Standardized: What would be the average outcome for Subgroup A if it had the same Covariate 1 distribution as Subgroup B

Answer

15 • (29/40) + 35 • (9/40) + 50 • (2/40) = 21.2

Smoking and mortality (Cochran, 1968)

Table 3

Adjusted Death Rates Using 3 Age Groups

Smoking Group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Covariates and outcomes

Definition (Predetermined Covariates)

Variable *X* is predetermined with respect to the treatment *T* if for each individual *i*, $X_{0i} = X_{1i}$, i.e. the value of **$\boldsymbol{X}_{\mathrm{i}}$ does not depend on the value of $\boldsymbol{T}_{\mathrm{i}}$**. Such characteristics are called *covariates*.

- Does not imply that *X* and *T* are independent.
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary.

Definition (Outcomes)

Those variables, *Y*, that are (possibly) not predetermined are called outcomes (for some individual *i*, $Y_{0i} \neq Y_{1i}$).

In general, one should not condition on outcomes, because this may induce bias.

Adjustment for observables in observational studies (Conditioning)

1. Subclassification

2. Matching

3. Regression

4. Propensity Score Methods (next time)

Identification under selection on observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp T|X$ *(selection on observables)*

   Potential outcomes (Y_1, Y_0) are independent of treatment variable (T) given we control for observables (X)

2. $0 < \Pr(T = 1|X) < 1$ *with probability one (common support)*

   Support is essentially the overlap between values of X for the comparison groups (defined by T=1 or 0)

Identification Result

Given selection on observables we have

$$E[Y_1 - Y_0|X] = E[Y_1 - Y_0|X, T=1]$$

$$= E[Y|X, T=1] - E[Y|X, T=0]$$

$$\alpha_{ATE} = E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X]dP(X)$$

Given we control for observable characteristics X, we can move forward with defining the average treatment effect as a difference in mean outcomes between treated and untreated individuals

Therefore, under the common support condition:

$$= \int \left(E[Y|X, T=1] - E[Y|X, T=0]\right)dP(X)$$

Identification under selection on observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp T|X$ *(selection on observables)*
2. $0 < \Pr(T = 1|X) < 1$ *with probability one (common support)*

Identification Result

Similarly,

$$\begin{aligned}\alpha_{ATET} &= E\left[Y_1 - Y_0 \middle| T = 1\right] \\ &= \int \left(E\left[Y \middle| X, T = 1\right] - E\left[Y \middle| X, T = 0\right] \right) dP\left(X \middle| T = 1\right)\end{aligned}$$

To identify α_{ATET} the selection on observables and common support conditions can be related to:

- $Y_0 \perp\!\!\!\perp T|X$
- $\Pr(T = 1|X) < 1$ *(with* $\Pr(T = 1|X) > 0$)

1. Subclassification estimator

The identification result is:

$$\alpha_{ATE} = \int \left(E\left[Y \middle| X, T=1 \right] - E\left[Y \middle| X, T=0 \right] \right) dP\left(X \right)$$

$$\alpha_{ATET} = \int \left(E\left[Y \middle| X, T=1 \right] - E\left[Y \middle| X, T=0 \right] \right) dP\left(X \middle| T=1 \right)$$

Assume *X* takes on *K* different cells {X^1, …, X^k, …, X^K}. Then, the analogy principle suggests the following estimators:

$$\hat{\alpha}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N^k}{N} \right); \quad \hat{\alpha}_{ATET} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k \right) \cdot \left(\frac{N_1^k}{N_1} \right)$$

- N^k is # of obs. and N_1^k is # of treated obs. in cell *k*
- $\bar{Y}_1^k$ is mean outcome for the treated in cell *k*
- $\bar{Y}_0^k$ is mean outcome for the untreated in cell *k*

1. Subclassification by age ($K = 2$)

	Death Rate	Death Rate		#	#
X_k	Smokers	Non-Smokers	Diff.	Smokers	Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Cells

Question

What is $\hat{\alpha}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$?

$$\hat{\alpha}_{ATE} = (28-24)\cdot\left(10/20\right) + (22-16)\cdot\left(10/20\right) = 5$$

*[(Difference in death rate between older smokers and nonsmokers) * (Proportion of smokers in the older group)] +*

*[(Difference in death rate between younger smokers and nonsmokers) * (Proportion of smokers in the younger group)] +*

1. Subclassification by age ($K = 2$)

	Death Rate	Death Rate		#	#
X_k	Smokers	Non-Smokers	Diff.	Smokers	Obs.
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question

What is $\hat{\alpha}_{ATET} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$ *?*

$$\hat{\alpha}_{ATET} = (28 - 24) \cdot (3/10) + (22 - 16) \cdot (7/10) = 5.4$$

1. Subclassification by age and gender (K = 4)

	Death Rate	Death Rate		#	#
X_k	Smokers	Non-Smokers	Diff.	Smokers	Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

Problem

What is $\hat{\alpha}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right)$ *?*

Not identified!

1. Subclassification by age and gender (K = 4)

X_k	Death Rate Smokers	Death Rate Non-Smokers	Diff.	# Smokers	# Obs.
Old, Male	28	22	4	3	7
Old, Female		24		0	3
Young, Male	21	16	5	3	4
Young, Female	23	17	6	4	6
Total				10	20

Problem

What is $\hat{\alpha}_{ATET} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$ *?*

$$\hat{\alpha}_{ATET} = 6 \cdot (3/10) + 5 \cdot (3/10) + 6 \cdot (4/10) = 5.7$$

1. Subclassification and the “curse of dimensionality”

- Subclassification becomes unfeasible with many covariates
- Assume we have *k* covariates and divide each of them into 3 coarse categories (e.g., age could be “young”, “middle age” or “old”, and income could be “low”, “medium” or “high”).
- The number of subclassification cells is 3^k. For $k = 10$, we obtain $3^{10} = 59049$
- Many cells may contain only treated or untreated observations, so we cannot use subclassification
- Subclassification is also problematic if the cells are “too coarse”. But using “finer” cell worsens the curse of dimensionality problem: e.g., using 10 variables and 5 categories for each variable we obtain $5^{10} = 9765625$

2. Matching

An alternative way to estimate α_{ATET} is by "imputing" the missing potential outcome of each treated unit using the observed outcome from the "closest" untreated unit:

$$\hat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{\{i|T_i=1\}} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the **closest** value to Y_i among the untreated observations.

We can also use the average for M closest matches:

$$\hat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{T_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)} \right) \right\}$$

Works well when we can find good matches for each treated unit, so M is usually small (typically, M = 1 or M = 2).

2. Matching

We can also use matching to estimate α_{ATE}. In that case, we match in both directions:

1. If observation *i* is treated, we impute Y_{0i} using untreated matches, $\{Y_{j_1(i)}, Y_{j_2(i)}, \ldots Y_{j_M(i)}\}$
2. If observation *i* is untreated, we impute Y_{1i} using treated matches, $\{Y_{j_1(i)}, Y_{j_2(i)}, \ldots Y_{j_M(i)}\}$

The estimator is:

$$\hat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(2T_i - 1\right)\left\{Y_i - \left(\frac{1}{M}\sum_{m=1}^{M} Y_{j_m(i)}\right)\right\}$$

Matching: example with single *X*

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	T_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question

What is $\hat{\alpha}_{ATET} = \frac{1}{N_i}\sum_{T_i=1}\left(Y_i - Y_{j(i)}\right)$ *?*

Match and plug in

Matching: example with single *X*

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question

What is $\hat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{D_i=1}\left(Y_i - Y_{j(i)}\right)$*?*

$\hat{\alpha}_{ATET}$ = 1/3 • (6 – 9) + 1/3 • (1 – 0) + 1/3 • (0 – 9) = -3.7

2. Matching: distance metric

When the vector of matching covariates,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix},$$

has more than one dimension ($k > 1$) we need to define a **distance metric** to measure "closeness". The usual **Euclidean distance** is:

$$\begin{aligned} \left\| X_i - X_j \right\| &= \sqrt{\left(X_i - X_j\right)' \left(X_i - X_j\right)} \\ &= \sqrt{\sum_{n=1}^{k} \left(X_{ni} - X_{nj}\right)^2}. \end{aligned}$$

→ The Euclidean distance is not invariant to changes in the scale of the *X*'s

→ For this reason, we often use alternative distances that are invariant to changes in scale

Matching: distance metric

A commonly used distance is the **normalized Euclidean distance:**

$$\left\| X_i - X_j \right\| = \sqrt{\left(X_i - X_j\right)' \hat{V}^{-1} \left(X_i - X_j\right)}$$

where

$$\hat{V} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_k^2 \end{pmatrix}$$

Notice that, the normalized Euclidean distance is equal to:

$$\left\| X_i - X_j \right\| = \sqrt{\sum_{n=1}^{k} \frac{\left(X_{ni} - X_{nj}\right)^2}{\hat{\sigma}_n^2}}.$$

→ Changes in the scale of X_{ni} affect also $\hat{\sigma}_n$, and the normalized Euclidean distance does not change

2. Matching: distance metric

Another popular scale-invariant distance is the **Mahalanobis distance:**

$$\left\| X_i - X_j \right\| = \sqrt{\left(\left(X_i - X_j \right)' \hat{\Sigma}_X^{-1} \left(X_i - X_j \right) \right)},$$

where $\hat{\Sigma}_X$ is the sample variance-covariance matrix of *X*.

We can also define arbitrary distances:

$$\left\| X_i - X_j \right\| = \sqrt{\sum_{n=1}^{k} \omega_n \cdot \left(X_{ni} - X_{nj} \right)^2}$$

(with all $\omega_n \geq 0$) so that we assign large ω_n‘s to those covariates that we want to match particularly well.

2. Matching and the curse of dimensionality

- Matching discrepancies $\left\| X_i - X_{j(i)} \right\|$ tend to increase with *k*, the dimension of *X*
- Matching discrepancies converge to zero as sample size increases. But they converge very slowly if *k* is large
- Mathematically, it can be shown that $\left\| X_i - X_{j(i)} \right\|$ converges to zero at the same rate as $\dfrac{1}{N^{1/k}}$
- It is difficult to find good matches in large dimensions: you need many observations if *k* is large

2. Matching: bias problem

We hope that we can apply a Central Limit Theorem and

$$\sqrt{N_1}\left(\hat{\alpha}_{ATET}-\alpha_{ATET}\right)$$

converges to a normal distribution with zero mean. However,

$$E\left[\sqrt{N_1}\left(\hat{\alpha}_{ATET}-\alpha_{ATET}\right)\right]=E\left[\sqrt{N_1}\left(\mu_0\left(X_i\right)-\mu_0\left(X_{j(i)}\right)\right)\middle|T=1\right].$$

Now, if *k* is large:

→ The difference between X_i *and* $X_{j(i)}$ converges to zero very slowly

→ The difference $\mu_0(X_i) - \mu_0(X_{j(i)})$ converges to zero very slowly

→ $E\left[\sqrt{N_1}\left(\mu_0\left(X_i\right)-\mu_0\left(X_{j(i)}\right)\right)\middle|T=1\right]$may not converge to zero!

→ $E\left[\sqrt{N_1}\left(\hat{\alpha}_{ATET}-\alpha_{ATET}\right)\right]$ may not converge to zero

→ **<u>Bias is often an issue when we match in many dimensions</u>**

2. Matching: Three solutions to the bias problem

The bias of the matching estimator is caused by large matching discrepancies $\left\| X_i - X_{j(i)} \right\|$. However:

1. The matching discrepancies are observed. We can always check in the data how well we are matching the covariates.

2. For $\hat{\alpha}_{ATET}$ we can always make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making N_0 large).

3. If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques.

4. Partial solution: Propensity score methods (to come).

2. Matching with bias correction

Abadie-Imbens

Each treated observation contributes

$$\mu_0(X_i)\text{-}\ \mu_0(X_{j(i)})$$

to the bias.

Bias corrected matching:

$$\hat{\alpha}_{ATET}^{BC} = \frac{1}{N_1}\sum_{T_i=1}\left(\left(Y_i - Y_{j(i)}\right) - \left(\hat{\mu}_0\left(X_i\right) - \hat{\mu}_0\left(X_{j(i)}\right)\right).\right)$$

where $\hat{\mu}_0\left(x\right)$ is an estimate of $E[Y|X = x,\ T = 0]$ (e.g., OLS).

Under some conditions, the bias correction eliminate the bias of the matching estimator without affecting the variance.

Bias adjustment in matched data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\alpha}_{ATET} = (10 - 8)/3 + (4 - 1)/3 + (10 - 9)/3 = 2$$

For the bias correction, estimate

$$\hat{\alpha}^{BC}_{ATET} = (\,(10 - 8) - (\ \hat{\mu}_0(3) - \hat{\mu}_0(4))\,)/3$$

$$+ (\,(4 - 1) - (\hat{\mu}_0(1) - \hat{\mu}_0\,(0))\,)/3$$

$$+ (\,(10 - 9) - (\,\hat{\mu}_0\,(10) - \hat{\mu}_0(8))\,)/3 = 1.33$$

Matching bias: implications for practice

Bias arises because of the effect of large matching discrepancies on $\mu_0(X_i) - \mu_0(X_{j(i)})$. To minimize matching discrepancies:

1. Use a small M (e.g., $M = 1$). Large values of M produce large matching discrepancies.
2. Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement (although greater variance).
3. Try to match covariates with a large effect on μ_0 (•) particularly well.

Matching estimators: large sample distribution (α_{ATET})

- Matching estimators have a Normal distribution in large samples (provided that the bias is small):

$$\sqrt{N_1}\left(\hat{\alpha}_{ATET}-\alpha_{ATET}\right)\xrightarrow{d}N\left(0,\sigma^2_{ATET}\right).$$

- For matching without replacement, the “usual” variance estimator,

$$\hat{\sigma}^2_{ATET}=\frac{1}{N_1}\sum_{T_i=1}\left(Y_i-\frac{1}{M}\sum_{m=1}^{M}Y_{j_m(i)}-\hat{\alpha}_{ATET}\right)^2,$$

is valid.

Matching estimators: large sample distribution (α_{ATET})

- For matching with replacement

$$\hat{\sigma}^2_{ATET} = \frac{1}{N_1}\sum_{T_i=1}\left(Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{j_m(i)} - \hat{\alpha}_{ATET}\right)^2 + \frac{1}{N_1}\sum_{T_i=0}\left(\frac{K_i(K_i-1)}{M^2}\right)\widehat{\text{var}}(\varepsilon_i \,|\, X_i, T_i = 0),$$

where K_i is the number of times observation *i* is used as a match.

- Can be estimated also by matching. For example take two observations with D_i=D_j=0 and X_i≅X_j, then

$$\widehat{\text{var}}(Y_i \,|\, X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of $\widehat{\text{var}}(\varepsilon_i \,|\, X_i, D_i = 0)$

- The bootstrap does not work!

Matching (with replacement) in Stata: nnmatch

- Basic usage:

  `nmatch` Y D X_1 X_2 `..., robust(1) pop`

- $M \neq 1$ (default is $M = 1$):

  `nmatch` Y D X_1 X_2 `..., m(2) robust(1) pop`

- ATT (default is ATE):

  `nmatch` Y D X_1 X_2 `..., tc(att) robust(1) pop`

- Mahalanbois metric (default is normalized Euclidean):

  `nmatch` Y D X_1 X_2 `..., metric(maha) robust(1) pop`

- Exact matches for some covariates:

  `nmatch` Y D X_1 X_2 `..., exact(`$\texttt{X}_{\texttt{exact}}$`) robust(1) pop`

- Bias correction (default is no bias correction):

  `nmatch` Y D X_1 X_2 `..., biasadj(bias) robust(1) pop`

Conditional Independence in Current Research

- *Complicated but cool emerging methods of reducing bias in matching designs*: “ Combining Matching and Synthetic Controls to Trade off Biases from Extrapolation and Interpolation” (2020) Maxwell Kellogg, Magne Mogstad, Guillaume Pouliot, Alexander Torgovitsky
- “Hidden in Plain Sight: Venture Growth with or without Venture Capital”, (2019) Christian Catalini, Jorge Guzman, Scott Stern
- *Again, very technical, but perhaps of interest to urban planning or transit-oriented folks:* “Inference via Low-Dimensional Couplings” (2018) Alessio Spantini, Daniele Bigoni, Youssef Marzouk