Week 3: Selection on Observables I
Matching and Subclassification
PUBL0050 Causal Inference

Julia de Romémont

Term 2 2023-24

UCL Departement of Political Science

Motivation

> *The planner of an observational study should always ask the question, "**How would the study be conducted if it were possible to do it by controlled experimentation?**"*
>
> *– Cochran, 1965*

- Randomization aids causal inference because in expectation it balances **observed** & **unobserved** confounders
- When we cannot randomize we can design studies to capture the central strength of randomized experiments:
  - have treatment & control groups that are as comparable as possible
  - i.e. we can try to **control for *observed*** covariates

This Lecture

Identification under Selection on Observables

Confounding and Post-Treatment Bias

Subclassification

Matching: Theory

Matching: Implementation

Identification under Selection on Observables

Intuition

Last week

- Randomization means that treatment assignment is independent of potential outcomes

This (and next) week

- Assume treatment is not randomized, but is independent of potential outcomes **so long as other factors are held fixed**

Intuition

We are assuming that among units with the same values for some covariate X (i.e. **conditional on** X), the treatment is "as good as randomly" assigned.

Identification Assumption

1. *Potential outcomes independent of D_i given X_i: $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$ ("selection on observables" or "conditional independence assumption")*
2. $0 < \Pr(D = 1|X) < 1$ *for all X (common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} E[Y_{1i} - Y_{0i}|X_i] &= E[Y_{1i} - Y_{0i}|X_i, D_i = 1] && \textbf{(CIA)} \\ &= E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1] \\ &= E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0] && \textbf{(CIA)} \\ &= E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] \end{aligned}$$

Implies that for any specific value for X_i, i.e. x_i, we can define the ***conditional*** *average treatment effect (δ_x):*

$$\delta_x \equiv E[Y_i|X_i = x, D_i = 1] - E[Y_i|X_i = x, D_i = 0]$$

Identification under Selection on Observables

Identification Assumption

1. *Potential outcomes independent of D_i given X_i: $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$ ("selection on observables" or "conditional independence assumption")*
2. $0 < \Pr(D = 1|X) < 1$ *for all X (common support)*

Identification Result

Therefore, under the common support condition and with a discrete X_i, we can calculate average effects of D_i on Y_i by taking weighted averages of δ_x:

$$\begin{aligned}
\hat{\tau}_{ATE} &= \sum_x \delta_x P(X_i = x) \\
\hat{\tau}_{ATT} &= \sum_x \delta_x P(X_i = x | D_i = 1) \\
\hat{\tau}_{ATC} &= \sum_x \delta_x P(X_i = x | D_i = 0)
\end{aligned}$$

i.e. where the weights are the distribution of X_i in the population ($\hat{\tau}_{ATE}$), treatment group ($\hat{\tau}_{ATT}$), and control group ($\hat{\tau}_{ATC}$).

Identification under Selection on Observables

This indentification assumption and result is common to all the methods we will study this week and next week.

- Subclassification **(today)**
- Matching **(today)**
- Regression **(next week)**

These differ in

a. how we condition on X_i and
b. how we weight δ_x.

Illustrative example

Does teacher training improve university applications?

Imagine that some school teachers take specialist training in how to prepare their students for university applications. Teachers select into the training program (i.e. they are not randomly assigned). You believe, however, that conditional on the type of school in which a teacher teachers, training is as good as random.

- Y_i: Number of students applying for top universities
- D_i: 1 if the teacher did the training, 0 otherwise
- X_i: Whether the teacher is at a state, private, or public school

Illustrative example

You collect some data and notice that teacher training is associated with teachers' school-types:

X_i, D_i joint distribution

	$D_i = 0$	$D_i = 1$
$X_i =$ State	0.30	0.05
$X_i =$ Private	0.15	0.15
$X_i =$ Public	0.05	0.30

You also notice that average student applications are associated with school-type and teacher training:

Mean outcomes

	$D_i = 0$	$D_i = 1$
$X_i =$ State	0	2
$X_i =$ Private	3	4
$X_i =$ Public	5	5

Illustrative example

X_i, D_i joint distribution		
	$D_i = 0$	$D_i = 1$
$X_i =$ State	0.30	0.05
$X_i =$ Private	0.15	0.15
$X_i =$ Public	0.05	0.30

Mean outcomes		
	$D_i = 0$	$D_i = 1$
$X_i =$ State	0	2
$X_i =$ Private	3	4
$X_i =$ Public	5	5

Given this information, calculate the difference in group means between teachers who did the extra training and those who did not:

$$\begin{aligned} \text{DIGM} &\equiv E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &= \frac{(0.05 \times 2 + 0.15 \times 4 + 0.3 \times 5)}{\frac{1}{2}} - \frac{(0.3 \times 0 + 0.15 \times 3 + 0.05 \times 5)}{\frac{1}{2}} \\ &= 3 \end{aligned}$$

Is the DIGM an unbiased estimator of the ATE?

No, we are assuming that the treatment is independent of potential outcomes **conditional on X**.

Illustrative example

- Selection on observables implies that the DIGM is an unbiased estimator for the ATE **within levels of** X_i.

So let's calculate those:

Mean outcomes

	$D_i = 0$	$D_i = 1$	δ_x
$X_i =$ State	0	2	2
$X_i =$ Private	3	4	1
$X_i =$ Public	5	5	0

- We can then summarize the effect of D_i on Y_i by taking **weighted averages** of δ_x
  - The weights are determined by our estimand of interest

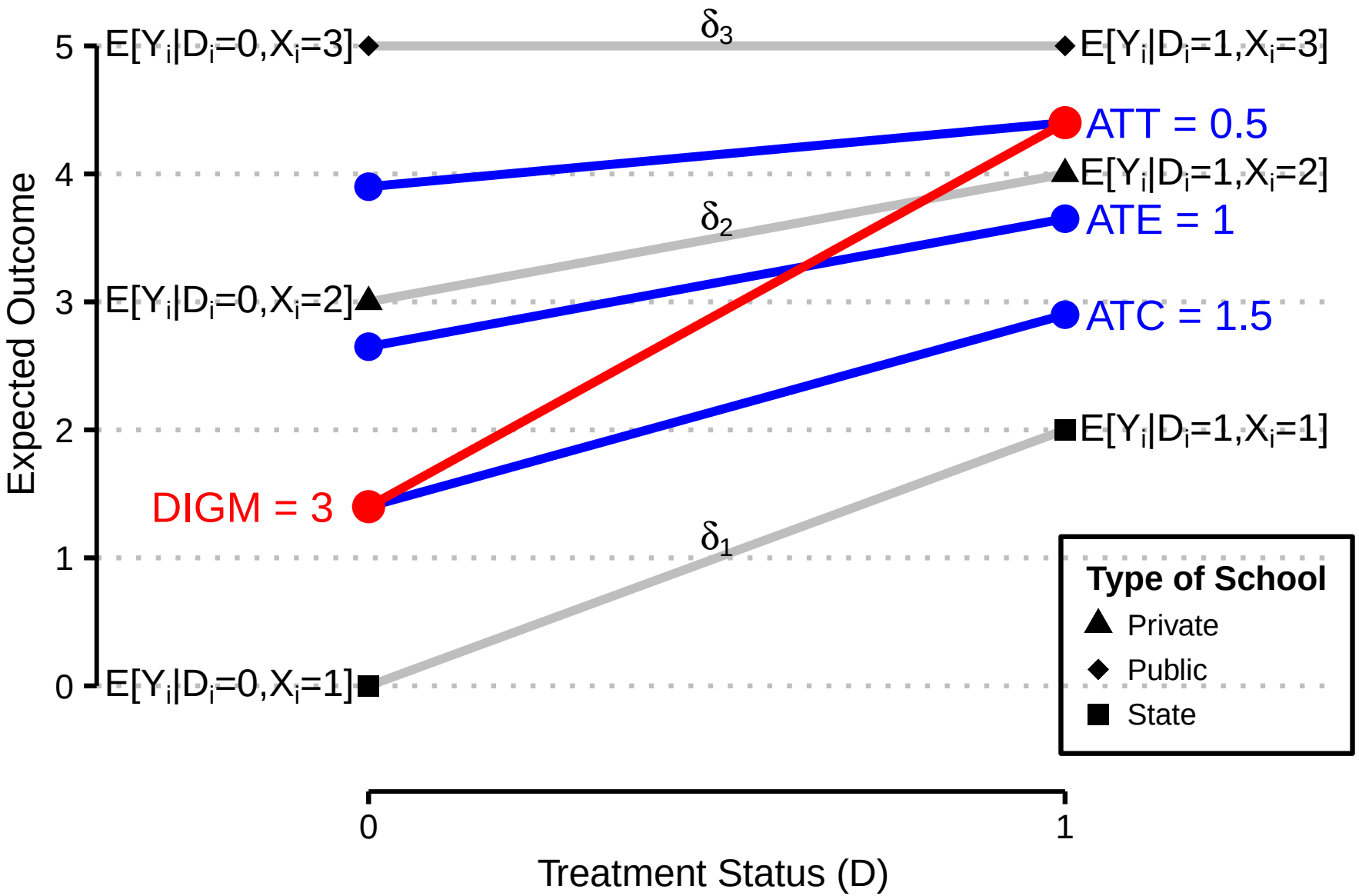
ATE vs ATT vs ATC

Distribution of X_i conditional on D_i			
	Control	Treatment	Population
$X_i =$ State	0.6	0.1	0.35
$X_i =$ Private	0.3	0.3	0.30
$X_i =$ Public	0.1	0.6	0.35

- ATE $\rightarrow$ weights δ_x by the distribution of X_i in the *population*
- ATT $\rightarrow$ weights δ_x by the distribution of X_i in the *treatment group*
- ATC $\rightarrow$ weights δ_x by the distribution of X_i in the *control group*

Common support implies that no weight is put on cells where there is 0 or 1 probability of treatment (because the δ_x is undefined)

Illustrative example



Illustrative example

ATE	ATT	ATC	DIGM
1	0.5	1.5	3

- Why is the DIGM bigger than all the $\hat{\tau}$ here?
  - ...Because of selection bias
- Why are ATE, ATT, and ATC the same in a randomized experiment?
  - ...Because the distribution of 'types of people' (according to a set of covariates X) between treatment and control group is **the same, in expectation**, due to randomisation
  - In a selection on observables design, **that is often not the case**

Confounding and Post-Treatment Bias

What should we select on?

Restate of the CIA:

Potential outcomes for control units are the same as for treated units, when those units have the same **covariate values** (X_i).

Question:

Which covariates make this assumption true?

Answer:

We do not know! This is an **untestable** assumption.

But, there are two ways in which it might fail:

1. **Selection bias** (confounding variable we have not included)
2. **Post-treatment bias** (included control is actually an outcome)

Selection bias $\equiv$ confounding

- Selection bias is just another name for **confounding**
- Confounding is the bias caused by common causes of the treatment and the outcome
- If we fail to account for **any** confounding variable Z_i that is related to both D_i and Y_i, then our identification assumption may be wrong
  - Though if our controls correlate with unobserved confounders, we might be OK
- In general, this is an **untestable assumption** though it is sometimes possible to provide indirect evidence
- More on this next week.

Identification under Selection on Observables is hard

> *We must remember that selection on observables is a large concession, which should not be made lightly. It is of far greater relevance than the technical discussion on the best way to condition on covariates. [...] The identification assumption for both OLS and matching is the same: selection on observables.*
>
> *– Sekhon, 2009*

- SOB research design depend entirely on the plausibility of treatment being *conditionally independent of potential outcomes*

Do not control for post-treatment variables

Does civic education increase voter turnout?

You are studying the effects of participating in a civic education programme on voter turnout. You also collect data on whether participants have high or low levels of political interest, where political interest is measured **after the education programme has been run**.

- Y_i is the outcome (voted $= 1$, not voted $= 0$)
- D_i is the treatment (participated $= 1$, did not participate $= 0$)
- Z_i is a **post-treatment** covariate (high interest $= 1$, low interest $= 0$)

We may wish to know the effects of education independent of political interest, so we might be tempted to control for political interest.

WE. SHOULD. NOT. DO. THIS.

Why?

Do not control for post-treatment variables

Notice first that every respondent has 2 potential Y_i outcomes, and 2 potential Z_i outcomes:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \qquad Z_i = \begin{cases} Z_{1i} & \text{if } D_i = 1 \\ Z_{0i} & \text{if } D_i = 0 \end{cases}$$

Consider the difference in mean outcome for those with high political interest:

$$\begin{aligned} \text{DIGM}_{Z_i=1} &= E[Y_i|Z_i = 1, D_i = 1] - E[Y_i|Z_i = 1, D_i = 0] \\ &= E[Y_{1i}|Z_{1i} = 1, D_i = 1] - E[Y_{0i}|Z_{0i} = 1, D_i = 0] \\ &= \underbrace{E[Y_{1i} - Y_{0i}|Z_{1i} = 1, D_i = 1]}_{\text{Causal effect}} + \\ &\qquad \underbrace{(E[Y_{0i}|Z_{1i} = 1, D_i = 1] - E[Y_{0i}|Z_{0i} = 1, D_i = 0])}_{\text{Selection bias}} \end{aligned}$$

These are not the same!

- $Z_{1i} = 1 \rightarrow$ High political interest **with** civic education
- $Z_{0i} = 1 \rightarrow$ High political interest **without** civic education

Do not control for post-treatment variables

- Intuition: introducing post-treatment variables means that you are, by design, not comparing similar units
- Post-treatment bias is a problem **even when the treatment is fully randomized** (i.e. experiments will not save you)
- Post-treatment bias is only not a problem if the treatment does not affect Z_i **(very difficult to establish in most settings)**
- Post-treatment bias can occur if
  - You control for a post-treatment variable
  - You control for a proxy variable that is measured after the treatment
  - You drop or select observations based on a post-treatment criterion

Overall lesson?

DO. NOT. CONDITION. ON. POST-TREATMENT. VARIABLES.

Subclassification

Running example

Do UN interventions Cause Peace?

Gilligan and Sergenti (2008) investigate whether UN peacekeeping operations have a causal effect on building sustainable peace after civil wars. They study 87 post-Cold-War conflicts, and evaluate whether peace lasts longer after conflict in 19 situations in which the UN had a peacekeeping mission compared to 68 situations where it did not.

- Y_i: Peace duration (measured in months)
- D_i: 1 if the UN intervened post-conflict, 0 otherwise
- $X_{1,i}$: Region of conflict (categorical)
- $X_{2,i}$: Democracy in the past (binary, based on polity)
- $X_{3,i}$: Ethnic Fractionalization (continuous)

Naive estimate

```
naive_diff <- mean(peace$dur[peace$UN == 1]) -
  mean(peace$dur[peace$UN == 0])
naive_diff
```

```
## [1] 74.4
```

Naive difference: peace lasted about 6 years longer, on average, in situations where the UN intervened.

Subclassification

- Subclassification is an estimation approach suitable for instances where we have categorical X_i variables (or where we make our X_i discrete)
- We already covered subclassification in the teachers example, but let's fix ideas here with our UN data.

Procedure:

1. Define subclasses
2. Calculate difference in mean outcome for treatment and control within each subclass
3. Calculate average treatment effects by taking weighted averages

Subclassification

Our subclasses here are the ten groups defined by the `region` and `democracy` variables:

Number of observations

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	10	3	21	7	8
Democracy	7	9	14	6	2

- For instance, in this sample there are:
  - 10 post-conflict instances in Eastern European countries that were formerly non-democracies
  - 14 post-conflict instances in Sub-saharan African countries that were formerly democracies
  - etc

Treated and control observations per subclass

How many treatment and control units do we have per subclass?

(Treated N, Control N)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	(5,5)	(2,1)	(4,17)	(0,7)	(1,7)
Democracy	(2,5)	(2,7)	(2,12)	(0,6)	(1,1)

- Is the common support assumption violated for any of these cells?

Difference in means per subclass

What is the conditional average treatment effect within each subclass?

δ_x (i.e. CATE)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	-29.2	144.0	27.9	NA	123.7
Democracy	66.8	5.1	49.0	NA	132.0

- Note that these are simply the difference in means estimates between treated and non-treated groups, within each subclass

Average Treatment Effects

(Treated N, Control N)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	(5,5)	(2,1)	(4,17)	(0,7)	(1,7)
Democracy	(2,5)	(2,7)	(2,12)	(0,6)	(1,1)

δ_x (i.e. CATE)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	-29.2	144.0	27.9	NA	123.7
Democracy	66.8	5.1	49.0	NA	132.0

$$
\begin{aligned}
ATT &= \sum_x \delta_x P(X_i = x | D_i = 1) \\
&= -29.2 \times {}^5/_{19} + 66.8 \times {}^2/_{19} + 144 \times {}^2/_{19} + 5.1 \times {}^2/_{19} \\
&\quad +27.9 \times {}^4/_{19} + 49 \times {}^2/_{19} + 123.7 \times {}^1/_{19} + 132 \times {}^1/_{19} \\
&= 39.53
\end{aligned}
$$

(i.e. weight by the proportion of **treated** observations in each cell)

Average Treatment Effects

(Treated N, Control N)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	(5,5)	(2,1)	(4,17)	(0,7)	(1,7)
Democracy	(2,5)	(2,7)	(2,12)	(0,6)	(1,1)

δ_x (i.e. CATE)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	-29.2	144.0	27.9	NA	123.7
Democracy	66.8	5.1	49.0	NA	132.0

$$
\begin{aligned}
ATE &= \sum_x \delta_x P(X_i = x) \\
&= -29.2 \times {}^{10}/_{74} + 66.8 \times {}^{7}/_{74} + 144 \times {}^{3}/_{74} + 5.1 \times {}^{9}/_{74} \\
&\quad +27.9 \times {}^{21}/_{74} + 49 \times {}^{14}/_{74} + 123.7 \times {}^{8}/_{74} + 132 \times {}^{2}/_{74} \\
&= 42.96
\end{aligned}
$$

(i.e. weight by the proportion of **all** observations in each cell)

Average Treatment Effects

(Treated N, Control N)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	(5,5)	(2,1)	(4,17)	(0,7)	(1,7)
Democracy	(2,5)	(2,7)	(2,12)	(0,6)	(1,1)

δ_x (i.e. CATE)

	E. Europe	L. America	SS Africa	Asia	N Africa/Middle East
Non-democracy	-29.2	144.0	27.9	NA	123.7
Democracy	66.8	5.1	49.0	NA	132.0

$$
\begin{aligned}
ATC &= \sum_x \delta_x P(X_i = x | D_i = 0) \\
&= -29.2 \times {}^5/_{55} + 66.8 \times {}^5/_{55} + 144 \times {}^1/_{55} + 5.1 \times {}^7/_{55} \\
&\quad +27.9 \times {}^{17}/_{55} + 49 \times {}^{12}/_{55} + 123.7 \times {}^7/_{55} + 132 \times {}^1/_{55} \\
&= 44.14
\end{aligned}
$$

(i.e. weight by the proportion of **control** observations in each cell)

Average Treatment Effects

$$\text{ATT} \rightarrow 39.53$$

$$\text{ATE} \rightarrow 42.96$$

$$\text{ATC} \rightarrow 44.14$$

These are all somewhat smaller than the raw DIGM (74.4).

Note that we are not really calculating ATC here:

- We cannot identify either of the δ_x for the Asian countries.
- ATC ends up being a somewhat odd quantity: the average treatment effect for the control observations that have overlap with treated observations.

Are $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | X_i$?

Subclassification is helpful for clarifying the CIA $\rightarrow$ we are assuming that treatment is "as good as random" within subclass.

Are we convinced by this assumption?

- Region and democratic history are probably not sufficient
  - What other X_i variables would it be important to condition upon?
- Subclassification is restricted to categorical X_i.
  - Not appropriate if key conditioning factors are continuous
  - e.g. Here we had to 'discrete-ize' the polity score

Is there a better way?

Matching: Theory

Recall the fundamental problem of causal inference:

$$Y_i = D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}$$

so

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

- One way of viewing this is as a **missing data problem**
  - i.e We observe half the potential outcomes for each unit, but not the other half
- One solution: **impute** the missing outcomes

  $\rightarrow$ This is what matching does

Matching

For each unit i, find the "closest" unit j with opposite treatment status and impute j's outcome as the unobserved potential outcome for i

$$\hat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

- where $Y_{j(i)}$ is the observed outcome for (untreated) unit j, the closest match to i
  - i.e. $X_{j(i)}$ is closest to X_i among the untreated observations.

It is also possible to use the average for the M closest matches:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - (\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}) \right\}$$

$\Rightarrow$ We could impute potential outcomes for control units and define the ATE/ATC equivalently.

Nearest Neighbour, Single X, $M = 1$, wth Replacement

NN 1:1 Matching

Country	D	EthFrac	Region	Y_{0i}	Y_{1i}
Liberia	1	83	SS Africa	?3	51
Sierra Leone	1	77	SS Africa	?11	35
Zaire	1	90	SS Africa	?3	23
Chad	0	83	SS Africa	3	
Senegal	0	72	SS Africa	11	
Niger	0	73	SS Africa	11	

What is the $\hat{\tau}_{ATT}$?

$$\begin{aligned}\hat{\tau}_{\text{ATT}} &= \frac{1}{N_1}\sum_{D_i=1}(Y_i - Y_{j(i)}) \\ &= (51-3)\times 1/3 + (35-11)\times 1/3 + (23-3)\times 1/3 \\ &= 30.7\end{aligned}$$

Nearest Neighbour, Single X, $M = 2$, wth Replacement

NN 2:1 Matching

Country	D	EthFrac	Region	Y_{0i}	Y_{1i}
Liberia	1	83	SS Africa	?7	51
Sierra Leone	1	77	SS Africa	?11	35
Zaire	1	90	SS Africa	?7	23
Chad	0	83	SS Africa	3	
Senegal	0	72	SS Africa	11	
Niger	0	73	SS Africa	11	

What is the $\hat{\tau}_{ATT}$?

$$\begin{aligned}
\hat{\tau}_{\text{ATT}} &= \frac{1}{N_1} \sum_{D_i=1} (Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)}) \\
&= (51 - 7) \times 1/3 + (35 - 11) \times 1/3 + (23 - 7) \times 1/3 \\
&= 28
\end{aligned}$$

More X variables

Commonly we will want to match on many X variables, not just one or two.

In our UN example, for instance, we might also include:

- Number of deaths in last war
- Duration of last war
- Ethnic fractionalization
- Military personnel
- Population size
- Mountains

Is this enough? What else? Are any of these post-treatment?

Matching distances

Adding more covariates creates a problem, however. We have to define how we measure whether two units are "close" to one another.

Which is "closer"?

- Treated case:
  - Haiti, with polity = -6, region = Latin America, and **ethfrac = 1**
- Control cases:
  - Panama, with polity = 8, region = Latin America, **ethfrac = 3**
  - Egypt, with polity = -7, region = N Africa, **ethfrac = 4**
  - El Salvador, with polity = -6, region = Latin America, **ethfrac = 26**

$\rightarrow$ We need a metric that takes 2 vectors of covariate values and projects them to a unidimensional scale

Matching distances

Exact matching

- Match j with i if j has identical covariates to i
- Rapidly breaks down with dimensionality of X, or with continuous X_i

Normalized Euclidian distance

- Scale distances on each variable by the inverse of sample variance
- Good with normally distributed X, not great with binary data

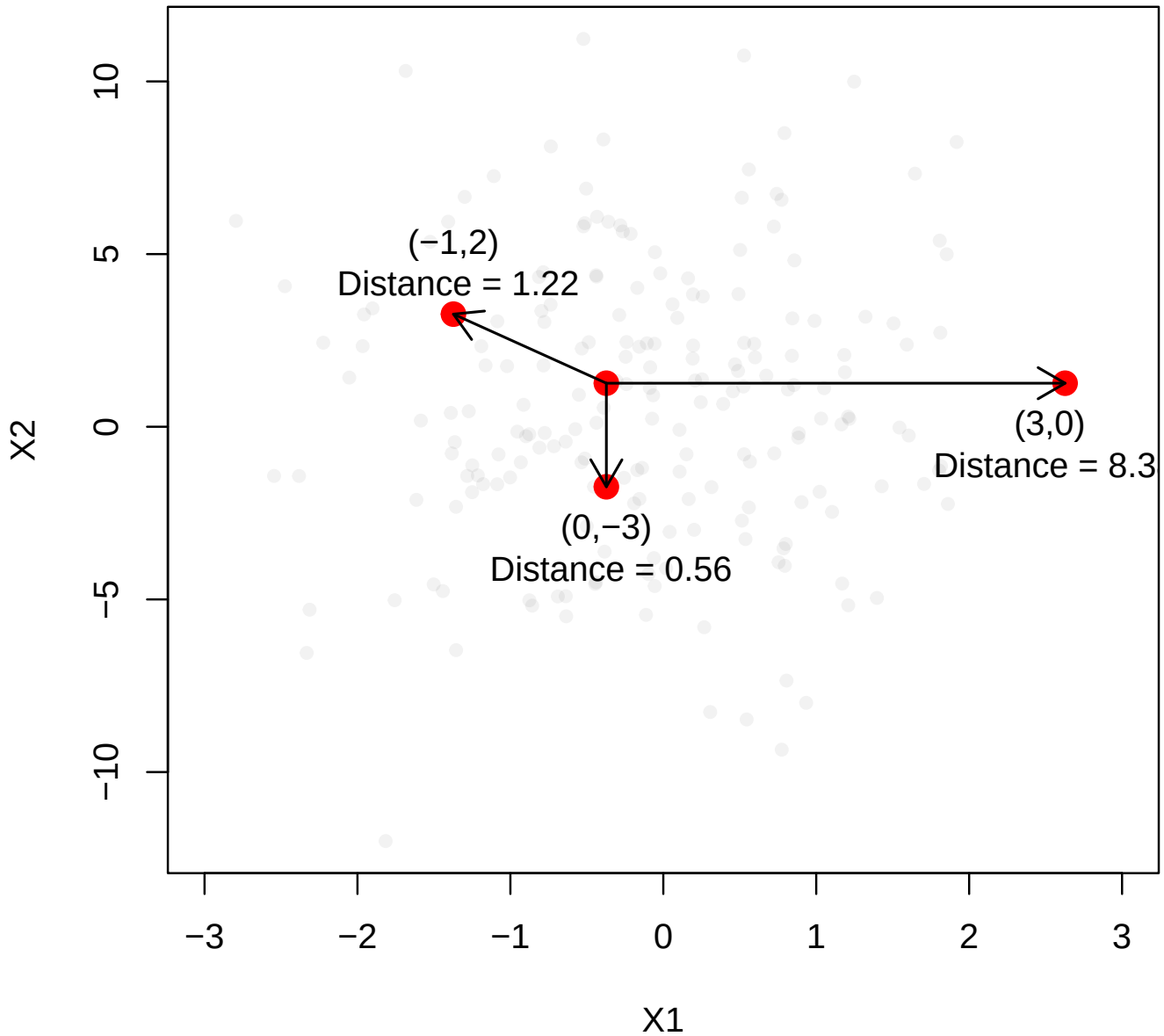
Mahalanobis distance

- Scale distances on each X by the inverse of the covariance matrix
- Good with normally distributed X, not great with binary data

Genetic matching

- Genetic matching aims directly to find the set of matches that minimize covariate imbalance across all variables

Mahalonobis distances



Other matching considerations

What size for M? 1-to-1? Many-to-1?

- Small M: decreased bias (2nd match always further than first)
- Large M: decreased variance (larger matched sample)

Matching with or without replacement?

- With replacement: decreased bias because some controls will be good matches for multiple treatment units
- But: replacement makes inference more complicated as matched controls are no longer independent (larger standard errors)

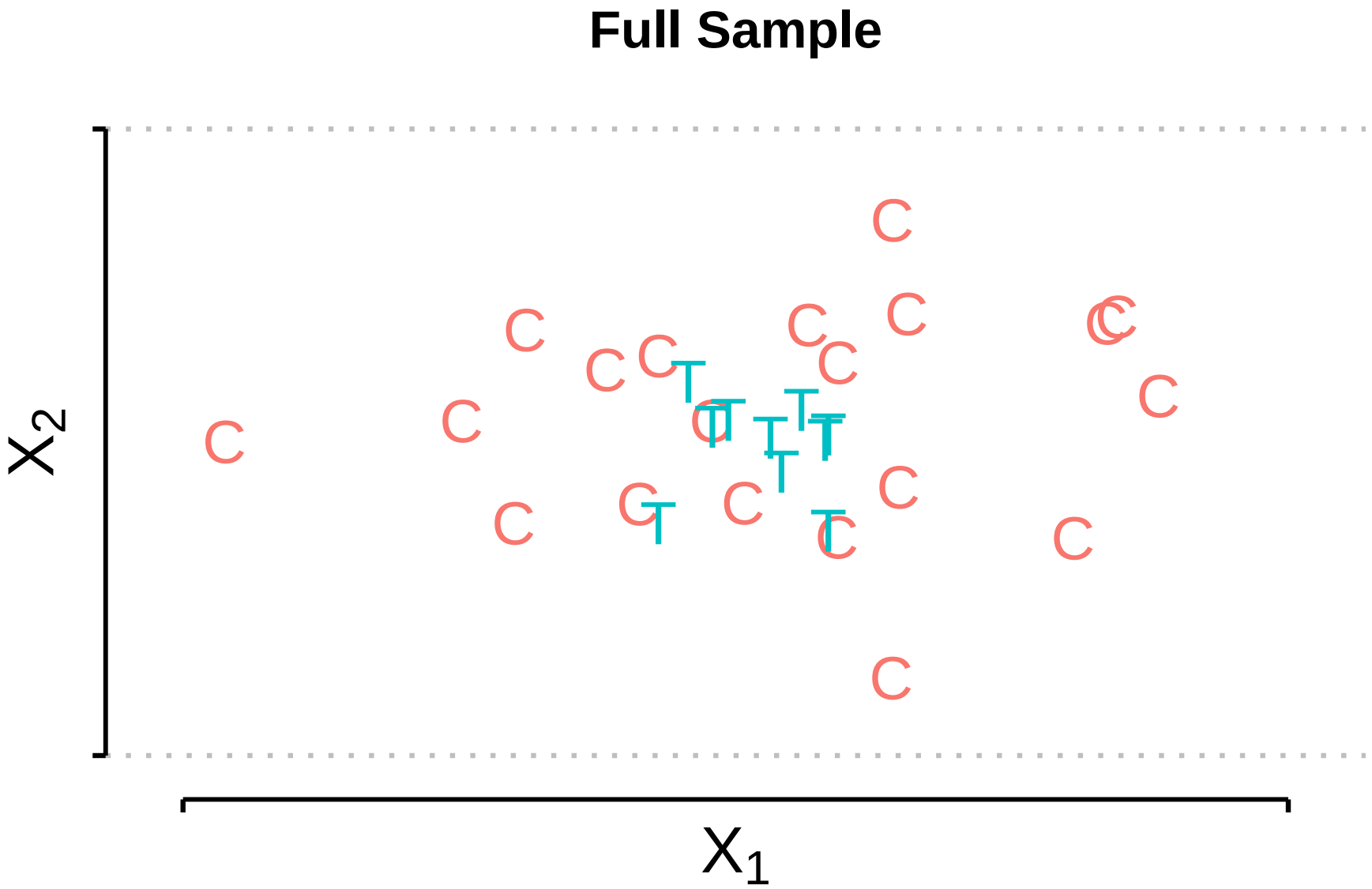
Breaking 'ties'

- When two controls are equally "close" to a treated unit
- Option 1: Select one at random (but: no unique answer)
- Option 2: Average the tied observed outcomes

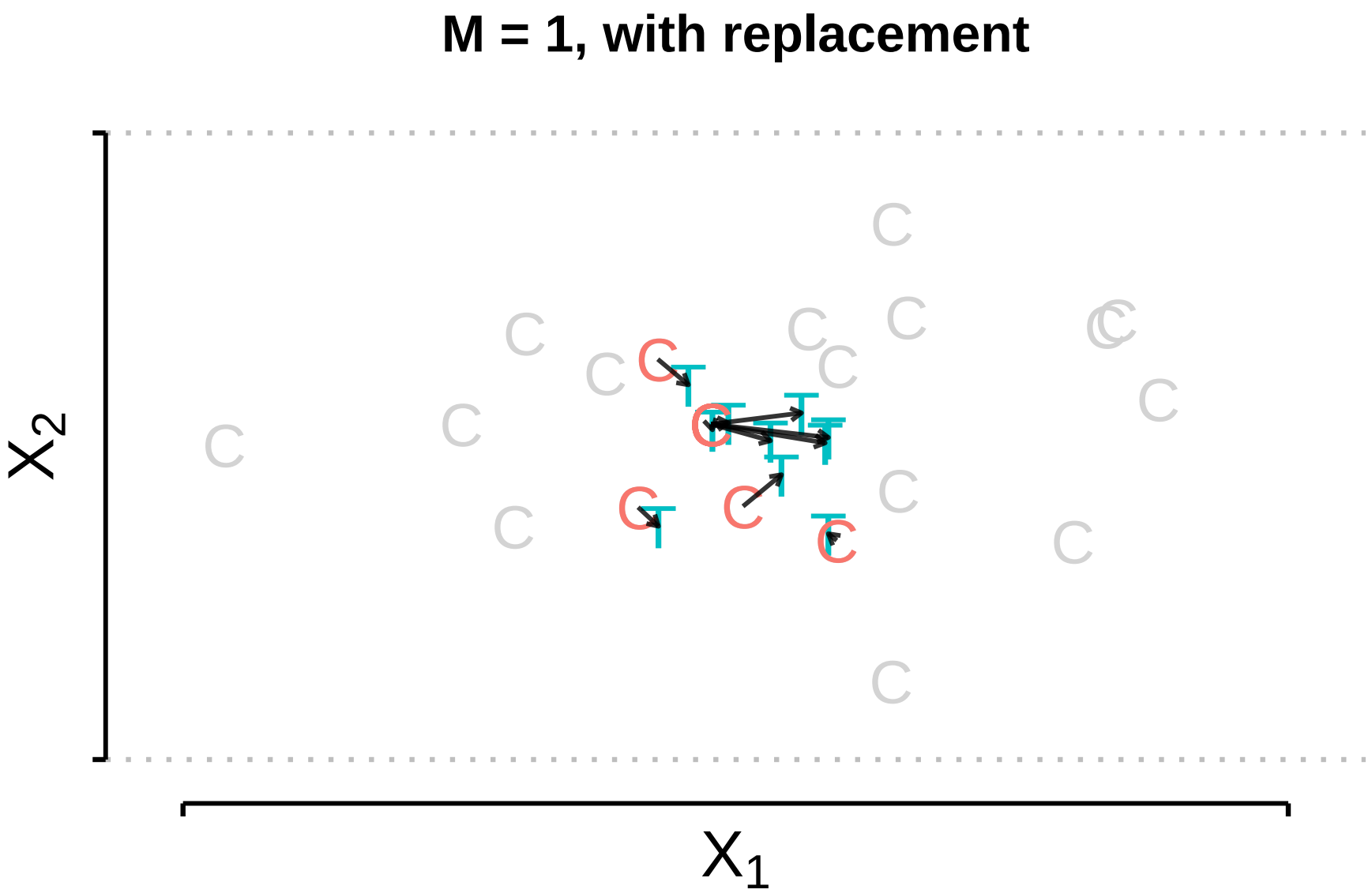
Which treatment effect? (ATE/ATC/ATT)

- Depends on substantive interest, also on available matches

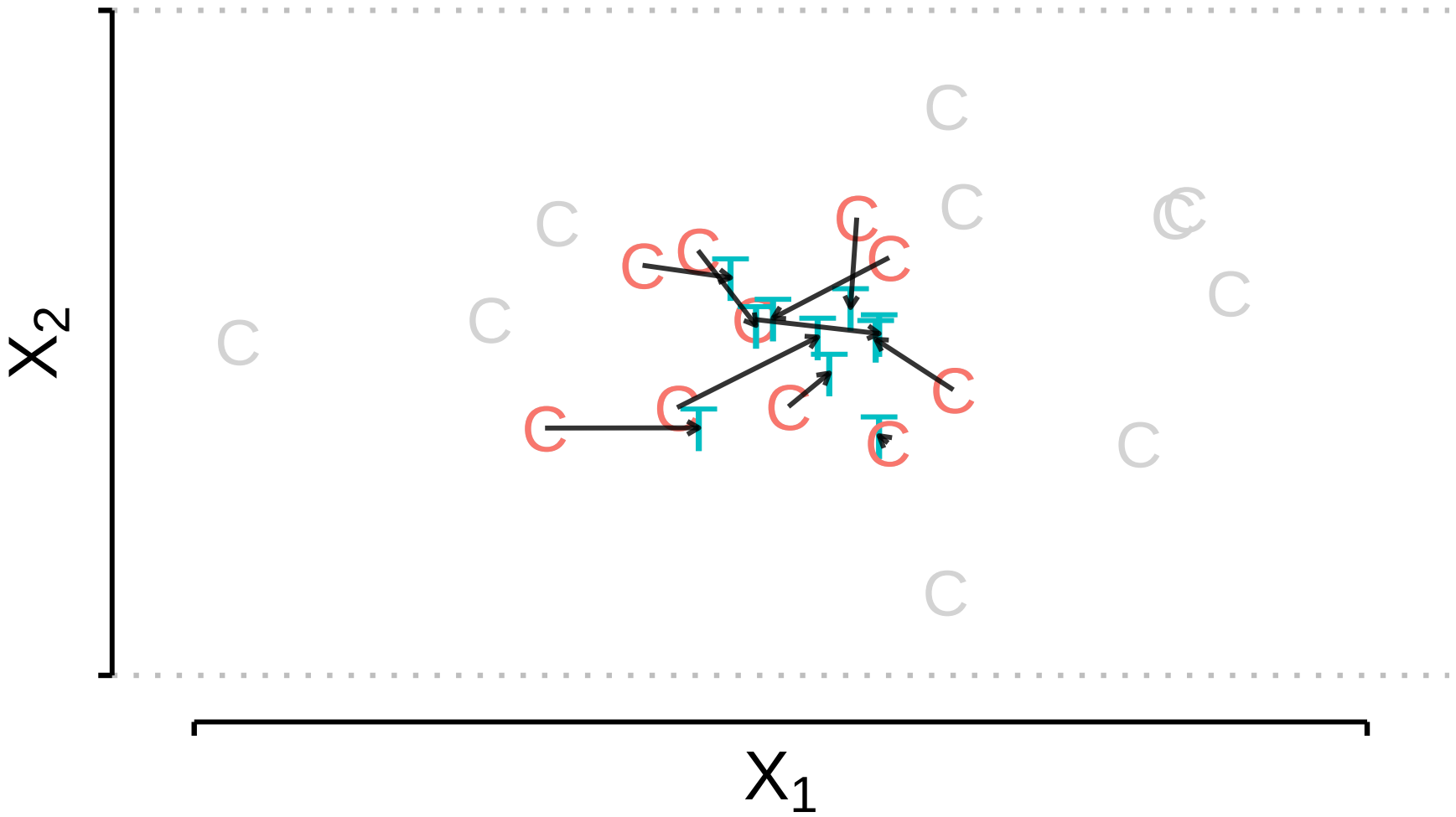
M and matching with/out replacement



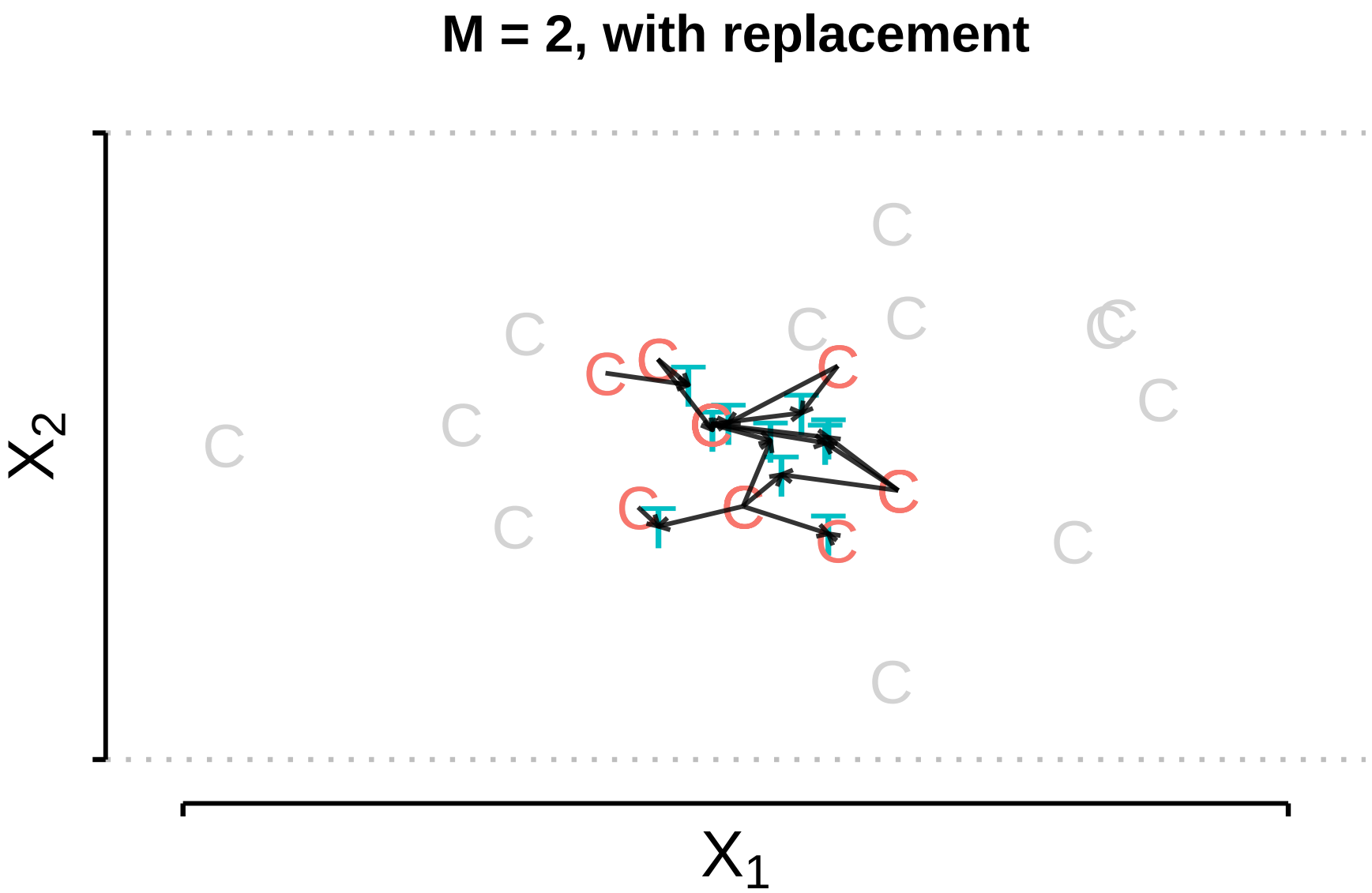
M and matching with/out replacement



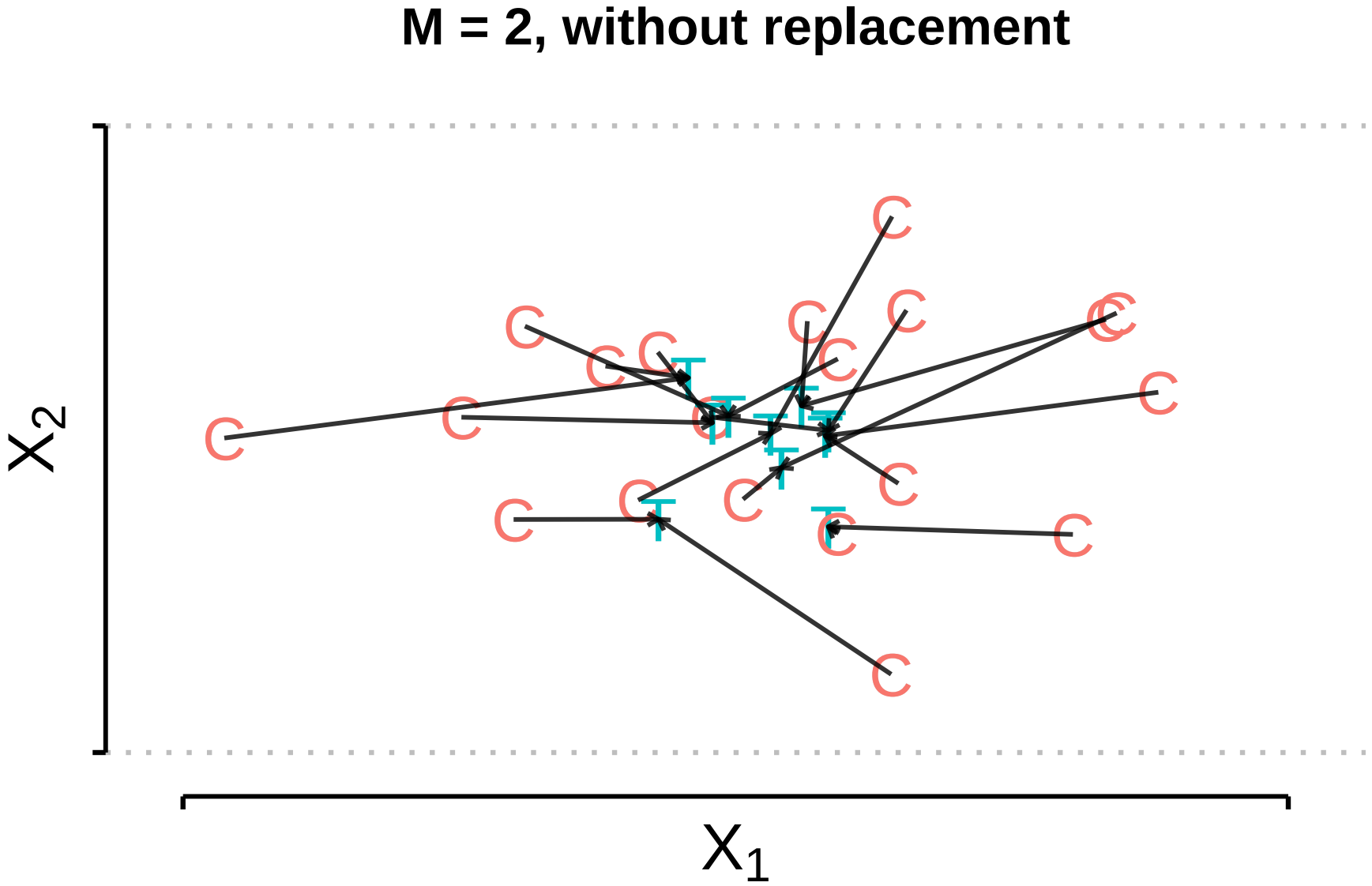
M and matching with/out replacement



M and matching with/out replacement



M and matching with/out replacement



Matching: Implementation

Matching in practice

```
library(MatchIt)

matched.out.att <- matchit(UN ~ lwdeaths + lwdurat + ethfrac + pop +
                             milper + bwplty2 + lmtnest + ssafrica  +
                             lamerica + eeurop,
                           data = peace,
                           method = "nearest", # nearest neighbour, default
                           distance = "mahalanobis", # can also be set manually
                           ratio = 1, # number of matches, default
                           replace = T, # with replacement
                           estimand = "ATT" # default
                           )
matched.out.att
```

```
## A matchit object
##  - method: 1:1 nearest neighbor matching with replacement
##  - distance: Mahalanobis
##  - number of obs.: 87 (original), 36 (matched)
##  - target estimand: ATT
##  - covariates: lwdeaths, lwdurat, ethfrac, pop, milper, bwplty2, lmtnest, ssaf
```

Matched data

```
matched.data.att <- match.data(matched.out.att)
matched.data.att[c(1,5,7),c("cname","dur","UN","weights")]
```

```
##        cname dur UN   weights
## 2      Haiti 143  1 1.0000000
## 9     Panama 169  0 0.8947368
## 12 Paraguay 177  0 1.7894737
```

```
# weights in treatment group (sum of weights = nobs in treatment group)
summary(matched.data.att$weights[matched.data.att$UN==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

```
# weights in control group (sum of weights = nobs in control group)
summary(matched.data.att$weights[matched.data.att$UN==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8947  0.8947  0.8947  1.0000  0.8947  1.7895
```

Matching weights

The weights are *key* here, as they ensure that treatment and control group are balanced with respect to the observable confounders.

- *With replacement:* $w_{j(i)} = \sum_{i=1}^{n_{i(j)}} \frac{1}{k_{i(j)}} \times \frac{n_j}{n_i}$
  - i are the observations for which j is a match
  - $k_{i(j)}$ is the total number of matches for i
  - n_i is the number of units with the same treatment status as i
  - n_j is the number of units with the same treatment status as j
- *Without replacement:* $w_{j(i)} = \frac{sp_{j(i)}}{1-sp_{j(i)}} \times \frac{n_j}{n_i}$
  - $sp_{j(i)}$ is the share of treated units in the matched pair or stratum j belongs to
- Sampling weights can also be incorporated by supplying the relevant variable name in the `s.weights =` option

Matching weights

Weights for control units in this example (ATT, 1:1, with replacement)

```
# Weight for control unit used twice
(1/1 + 1/1)*17/19
```

```
## [1] 1.789474
```

```
matched.data.att$weights[matched.data.att$cname=="Paraguay"]
```

```
## [1] 1.789474
```

```
# Weight for control unit used once
(1/1)*17/19
```

```
## [1] 0.8947368
```

```
matched.data.att$weights[matched.data.att$cname=="Panama"]
```

```
## [1] 0.8947368
```

Calculating the estimate

```
# Now estimate the treatment effect (ATT) with regression
match.att <- lm(dur ~ UN ,
                data = matched.data.att,
                weights = weights) ## !
# ATT
coef(match.att)[2]
```

```
##        UN
## 20.84211
```

Assessing balance

After matching, the distribution of X should be the same for treatment and control groups:

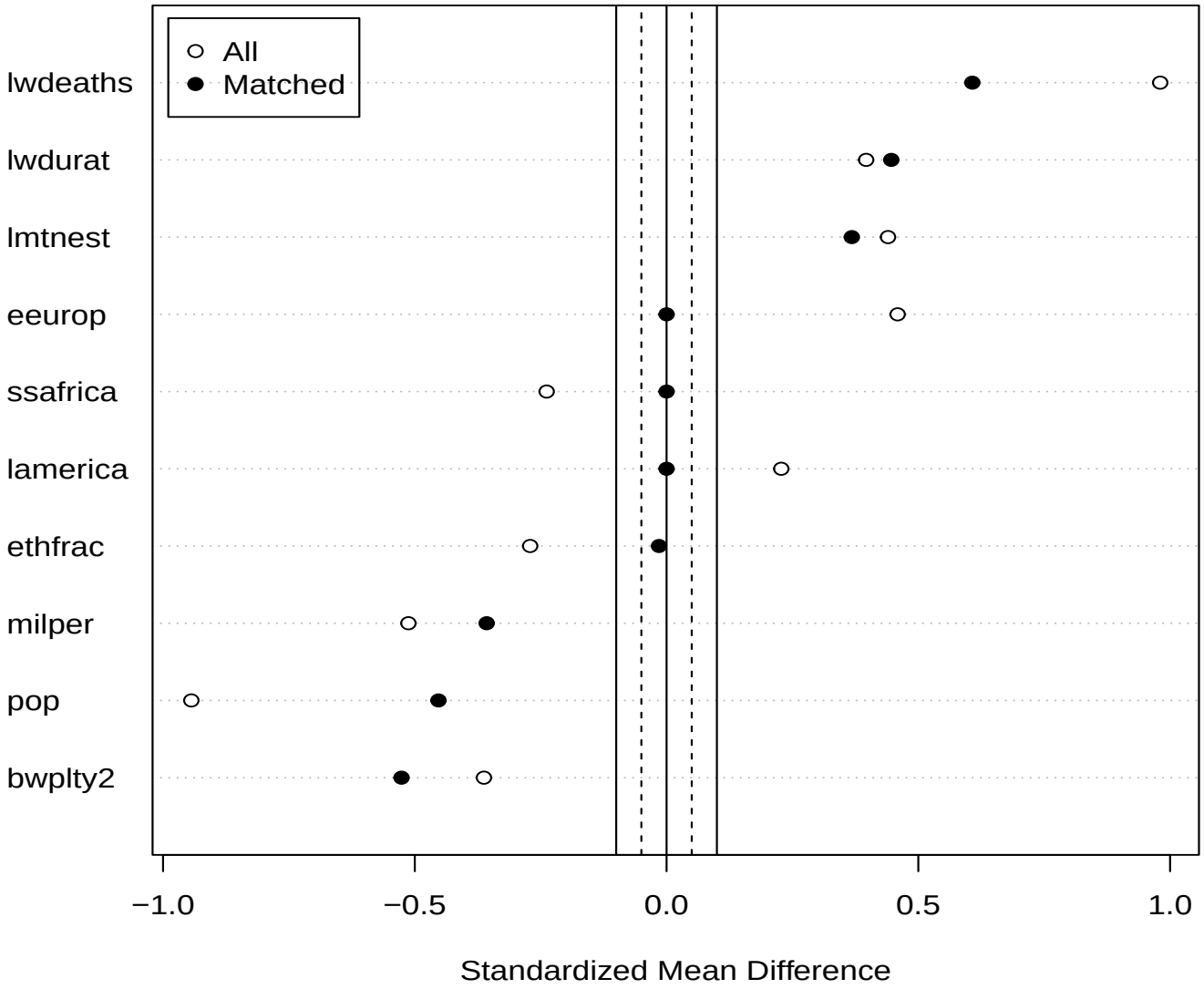
- Many papers will present tables of covariate means and p-values before and after matching as evidence of comparability
- Strictly speaking, p-value are not very informative, as they are sensitive to changes in the sample size
- Instead, it is useful to measure the **standardized bias** of a covariate before and after matching and compare:

$$\text{bias}_{X_i} = \frac{\bar{X}_t - \bar{X}_c}{\sigma_t}$$

where σ_t is the standard deviation of X in the full treated group.

Assessing balance

```
plot(summary(matched.out.att), abs = F,
     position = "topleft",var.order = "matched")
```



Genetic matching in practice

```
set.seed(123)
gen.matched.out.att <- matchit(UN ~ lwdeaths + lwdurat + ethfrac + pop +
                                 milper + bwplty2 + lmtnest + ssafrica  +
                                 lamerica + eeurop,
                               data = peace,
                               method = "genetic",
                               distance = "mahalanobis",
                               ratio = 1,
                               replace = T,
                               estimand = "ATT",
                               pop.size=1000)
gen.matched.out.att
```

```
## A matchit object
##  - method: 1:1 genetic matching with replacement
##  - distance: Mahalanobis
##  - number of obs.: 87 (original), 33 (matched)
##  - target estimand: ATT
##  - covariates: lwdeaths, lwdurat, ethfrac, pop, milper, bwplty2, lmtnest, ssaf
```

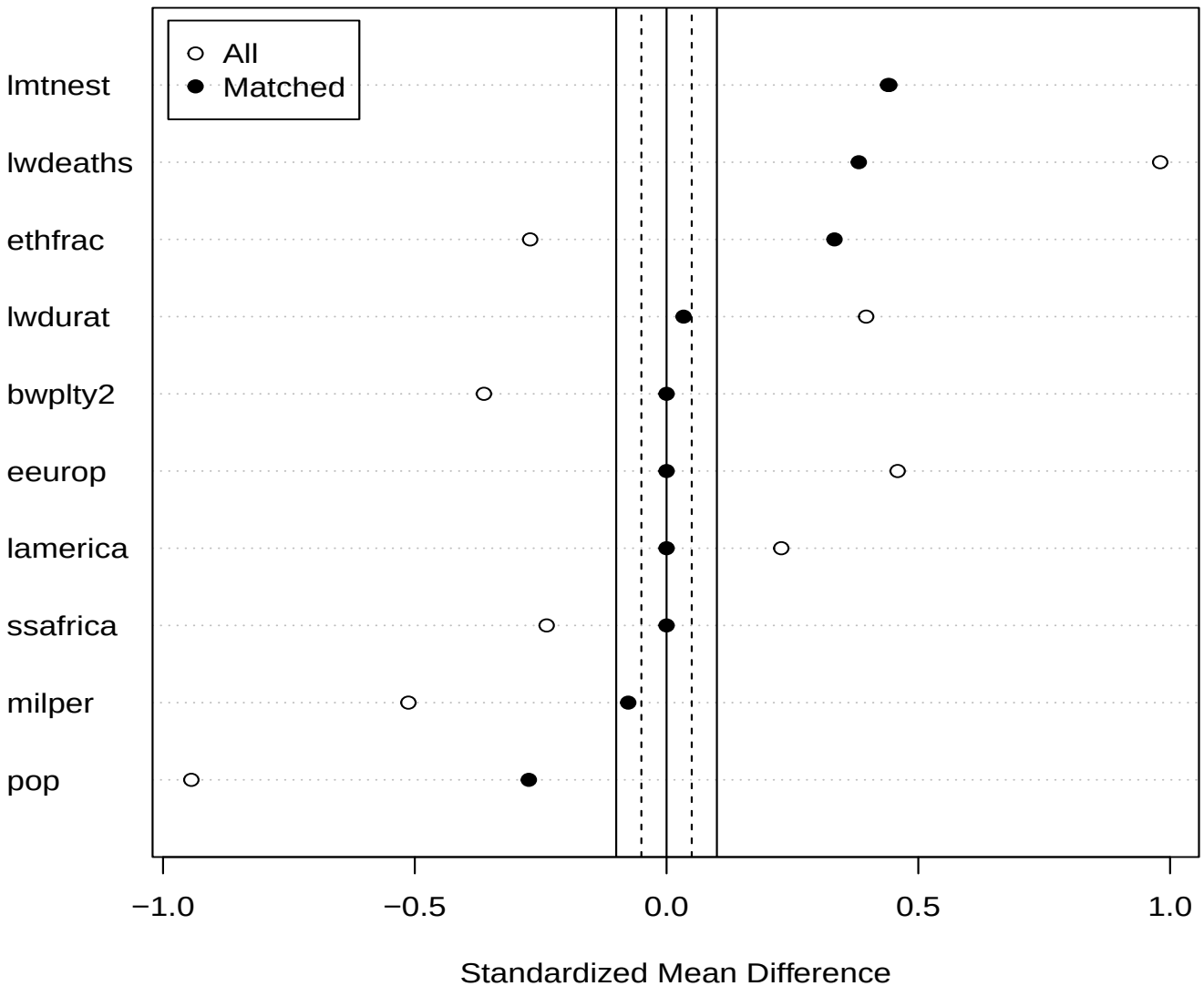
Calculating the estimate

```
# get matched data
gen.matched.data.att <- match.data(gen.matched.out.att)

# Now estimate with regression
gen.match.att <- lm(dur ~ UN,
                    data = gen.matched.data.att,
                    weights = weights)

# ATT
coef(gen.match.att)[2]
```

```
##        UN
## 35.36842
```

Assessing balance

```
plot(summary(gen.matched.out.att),abs = F,
     position = "topleft",var.order = "matched")
```



Consequences of matching decisions

ATT's from different matches

M	Replacement	Distance	ATT
1	Yes	euclidean	43.47
1	No	euclidean	51.90
2	Yes	euclidean	33.12
2	No	euclidean	44.55
1	Yes	mahalanobis	22.98
1	No	mahalanobis	21.04
2	Yes	mahalanobis	32.92
2	No	mahalanobis	37.13
1	Yes	genetic	50.57
1	No	genetic	44.90
2	Yes	genetic	48.59
2	No	genetic	28.50

Implication: Even using the same covariates, different matching criteria can lead to different outcomes! Particularly when N is small.

Consequences of matching decisions

One approach is to pick the matching procedure that results in the smallest standardized difference in means across all covariates:

Bias & ATT's from different matches

M	Replacement	Distance	ATT	Mean absolute bias
1	Yes	euclidean	43.47	0.36
1	No	euclidean	51.9	0.32
2	Yes	euclidean	33.12	0.4
2	No	euclidean	44.55	0.42
1	Yes	mahalanobis	22.98	0.28
1	No	mahalanobis	21.04	0.29
2	Yes	mahalanobis	32.92	0.26
2	No	mahalanobis	37.13	0.34
1	Yes	genetic	50.57	0.13
1	No	genetic	44.9	0.18
2	Yes	genetic	48.59	0.17
2	No	genetic	28.5	0.31

Matching advice

> *While extensive time and effort is put into the careful design of randomized experiments, relatively little effort is put into the corresponding 'design' of non-experimental studies.*
>
> *– Stuart, 2010*

Best practice is to design without access to outcome variables:

1. Look at data without outcome variables; design matching strategy
   - 1 to 1; many to 1; with/without replacement, etc
2. Test covariate balance; if unbalanced, go back to 1
3. Compare outcomes only after matching is completed.

Conclusion

- By assuming treatments are "as good as random" conditional on X, we can make causal claims from non-experimental data
- How convincing our causal claims are is entirely determined by how plausible this assumption seems in a given context
- We should condition on all potentially confounding variables
- We should not condition on any post-treatment variables
- Matching and subclassification are two approaches to conditioning