THE UNIVERSITY
of EDINBURGH

Methods for Causal Inference
Lecture 4: Rubin's potential outcomes framework

Ava Khamseh

School of Informatics

2024-2025

Last two lectures …

Language of probability: Variables, evens, samples space, probability law

Probability axioms, (conditional) total law of probability, independence, Bayes’ rule

Expected values, variance, correlation

Graphs

Today:

First of the two causal frameworks:

- **Potential Outcomes (due to Neyman-Rubin)**
- **Study our first causal question**

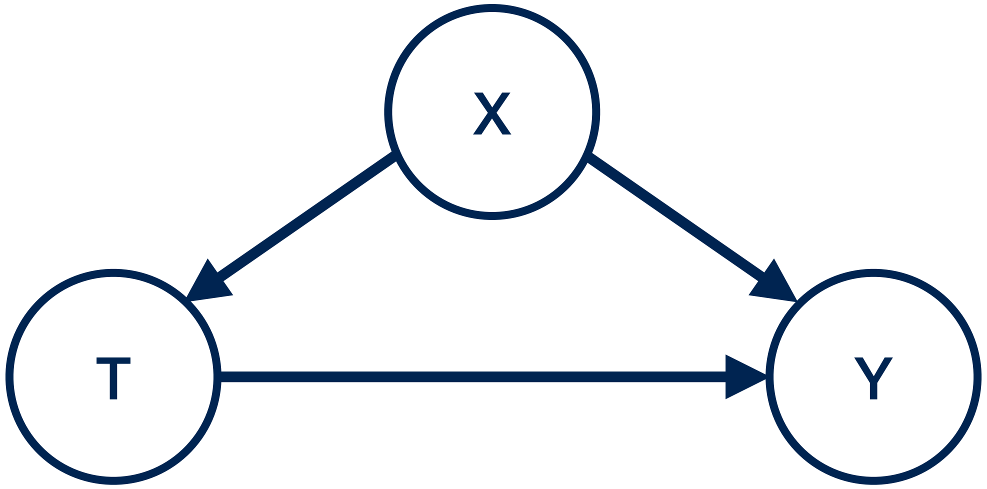
In order to estimate:

- Answer to a causal question
- Uncertainty on this answer (under model assumptions)

Two main Frameworks for causal identifiablity

- Potential outcomes framework (Neyman-Rubin):

- Requires a given treatment-outcome pair (known directionality)
- For causal estimation
- More familiar to biomedical researchers (this is changing ...)



- Structural causal models (Pearl):

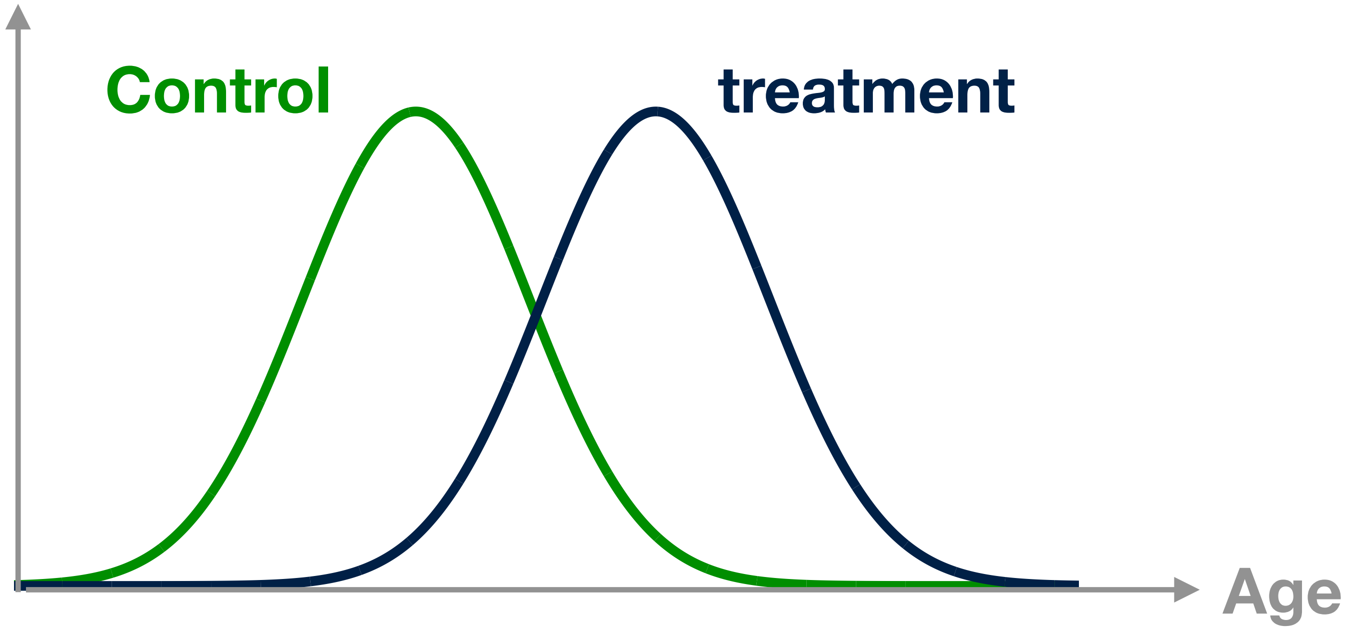
- Causal graphs
- Structural equations $x = f_x(\epsilon_x),\ t = f_t(x, \epsilon_t),\ y = f_y(x, t, \epsilon_y)$
- Algorithmic
- For causal estimation and discovery

Assumption: Independent noise terms: $\epsilon_x \perp\!\!\!\perp \epsilon_t \perp\!\!\!\perp \epsilon_y$

Extend the language of probability theory:
do-calculus

Observational data: What goes wrong?

$$p(x|t=1) \neq p(x|t=0)$$



$$\left(\int y_1(x)p(x|t=1)dx \;-\int y_0(x)p(x|t=0)dx\right) \neq \int \big(y_1(x)-y_0(x)\big)p(x)dx$$

Observational data: Stratification

- Measure outcome (success/failure), **within** each of the young/old groups **separately**
- Take weighted average by the probability of being young/old:

$$\mathbb{E}(\mathrm{Healed}|t=1) = \mathbb{E}(\mathrm{Healed}|t=1, \mathrm{young})p(\mathrm{young}) + \mathbb{E}(\mathrm{Healed}|t=1, \mathrm{old})p(\mathrm{old})$$

vs

$$\mathbb{E}(\mathrm{Healed}|t=0) = \mathbb{E}(\mathrm{Healed}|t=0, \mathrm{young})p(\mathrm{young}) + \mathbb{E}(\mathrm{Healed}|t=0, \mathrm{old})p(\mathrm{old})$$

Observational data: Stratification

- Measure outcome (success/failure), **within** each of the young/old groups **separately**
- Take weighted average by the probability of being young/old:

$$\mathbb{E}(\text{Healed}|t=1) = \mathbb{E}(\text{Healed}|t=1,\text{young})p(\text{young}) + \mathbb{E}(\text{Healed}|t=1,\text{old})p(\text{old})$$

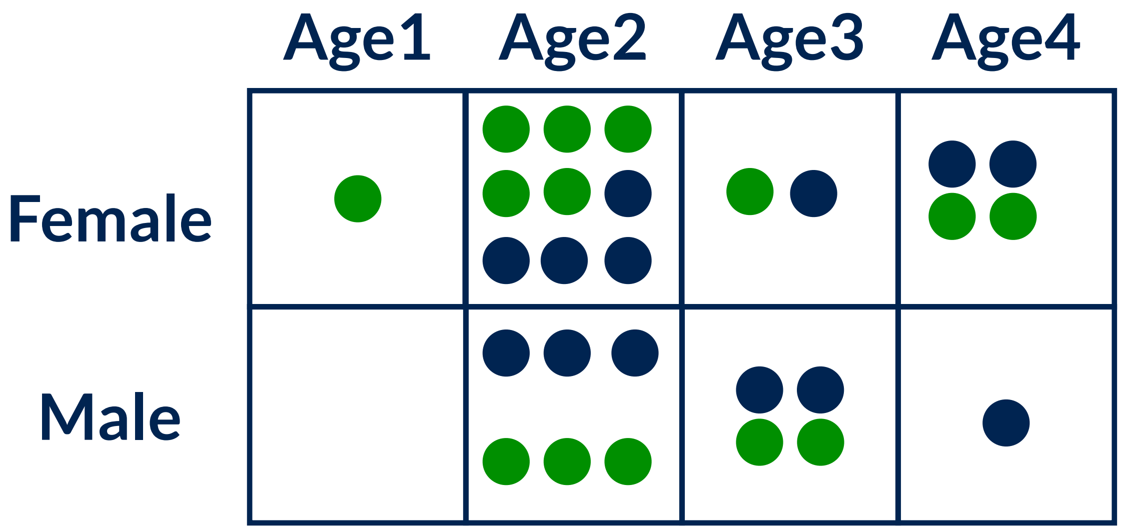
vs

$$\mathbb{E}(\text{Healed}|t=0) = \mathbb{E}(\text{Healed}|t=0,\text{young})p(\text{young}) + \mathbb{E}(\text{Healed}|t=0,\text{old})p(\text{old})$$

<u>Issues:</u> (i) All possible confounders need to be observed

(ii) Assume overlap between the two distributions (if there is no overlap, sample is not representative, e.g. performing the experiment only for old people),

(iii) Poor estimates as confounder dimensionality increases



⇒

Need specific causal effect estimation techniques

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

$y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes

$t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1

Observed outcomes: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)} = \begin{cases} y_0^{(i)} & \text{if } \ t^{(i)} = 0 \\ y_1^{(i)} & \text{if } \ t^{(i)} = 1 \end{cases}$$

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

$y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes

$t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1

Observed outcomes: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)} = \begin{cases} y_0^{(i)} & \text{if} \quad t^{(i)} = 0 \\ y_1^{(i)} & \text{if} \quad t^{(i)} = 1 \end{cases}$$

Counterfactual (missing) outcome "what would have happened if …"

$$y_{CF}^{(i)} = (1 - t^{(i)}) y_1^{(i)} + t^{(i)} y_0^{(i)} = \begin{cases} y_1^{(i)} & \text{if } t^{(i)} = 0 \\ y_0^{(i)} & \text{if } t^{(i)} = 1 \end{cases}$$

Potential Outcomes Framework (Rubin-Neyman)

Inverting previous relations, equivalently:

$$y_0^{(i)} = \begin{cases} y_{CF}^{(i)} & \text{if} \quad t^{(i)} = 1 \\ y_{obs}^{(i)} & \text{if} \quad t^{(i)} = 0 \end{cases}$$

$$y_1^{(i)} = \begin{cases} y_{CF}^{(i)} & \text{if} \quad t^{(i)} = 0 \\ y_{obs}^{(i)} & \text{if} \quad t^{(i)} = 1 \end{cases}$$

Knowing the potential outcomes is equivalent to knowing the observed and counterfactual outcomes

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

$y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes

$t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1

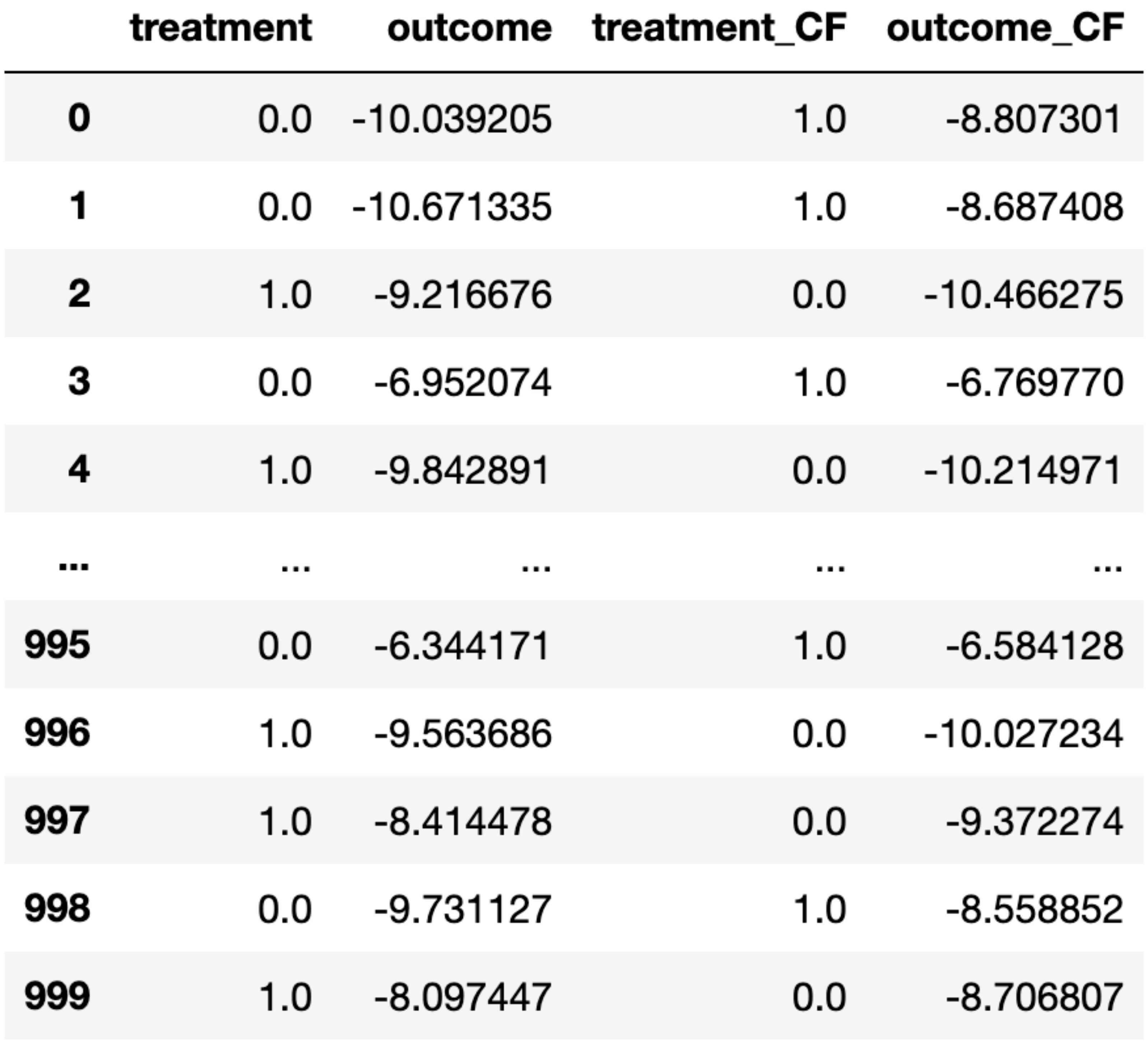
Observed outcomes: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)}$$

Individual treatment effect (causal): $\tau^{(i)} = y_1^{(i)} - y_0^{(i)}$

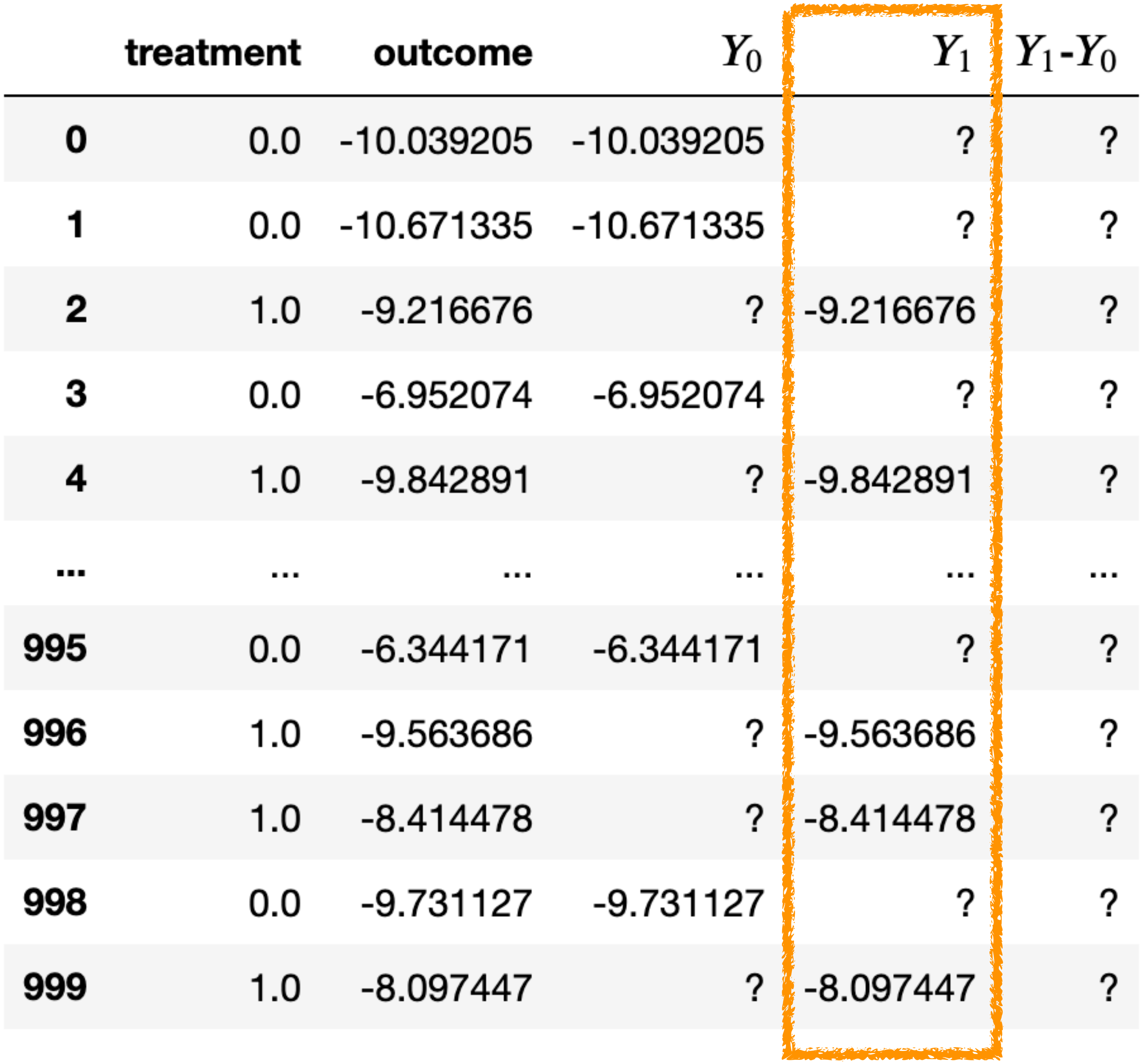
Average treatment effect (causal): $\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N} \sum_{i=0}^{N} \left(y_1^{(i)} - y_0^{(i)} \right)$

Example (Missing data interpretation)



	treatment	outcome	treatment_CF	outcome_CF
0	0.0	-10.039205	1.0	-8.807301
1	0.0	-10.671335	1.0	-8.687408
2	1.0	-9.216676	0.0	-10.466275
3	0.0	-6.952074	1.0	-6.769770
4	1.0	-9.842891	0.0	-10.214971
...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128
996	1.0	-9.563686	0.0	-10.027234
997	1.0	-8.414478	0.0	-9.372274
998	0.0	-9.731127	1.0	-8.558852
999	1.0	-8.097447	0.0	-8.706807

Example (Missing data interpretation)



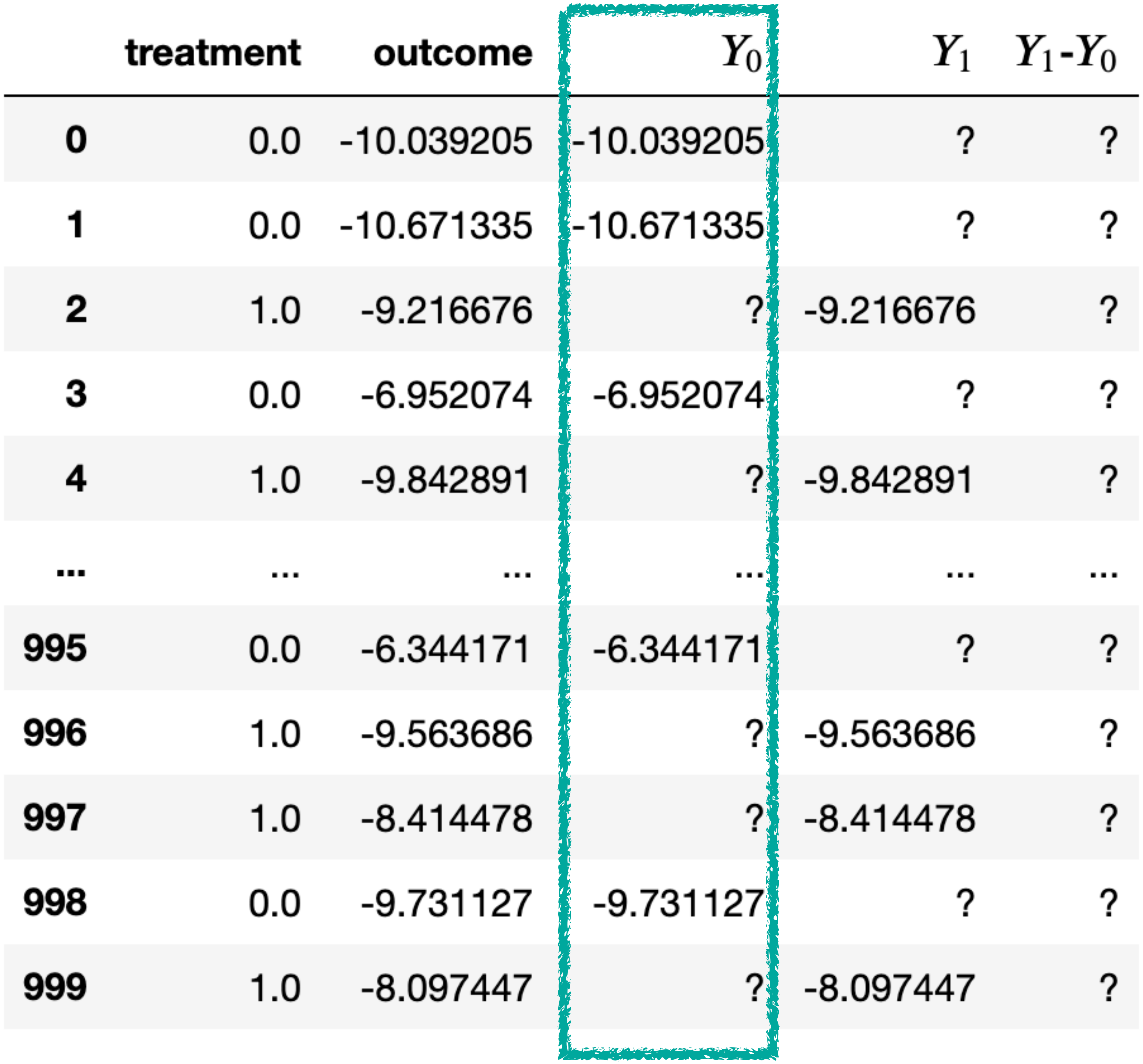
	treatment	outcome	Y_0	Y_1	Y_1-Y_0
0	0.0	-10.039205	-10.039205	?	?
1	0.0	-10.671335	-10.671335	?	?
2	1.0	-9.216676	?	-9.216676	?
3	0.0	-6.952074	-6.952074	?	?
4	1.0	-9.842891	?	-9.842891	?
...	...	...	...	...	...
995	0.0	-6.344171	-6.344171	?	?
996	1.0	-9.563686	?	-9.563686	?
997	1.0	-8.414478	?	-8.414478	?
998	0.0	-9.731127	-9.731127	?	?
999	1.0	-8.097447	?	-8.097447	?

What about the naive observational estimator?

$$\boxed{\mathbb{E}[Y|T=1]} - \mathbb{E}[Y|T=0]$$

-9.70

Example (Missing data interpretation)



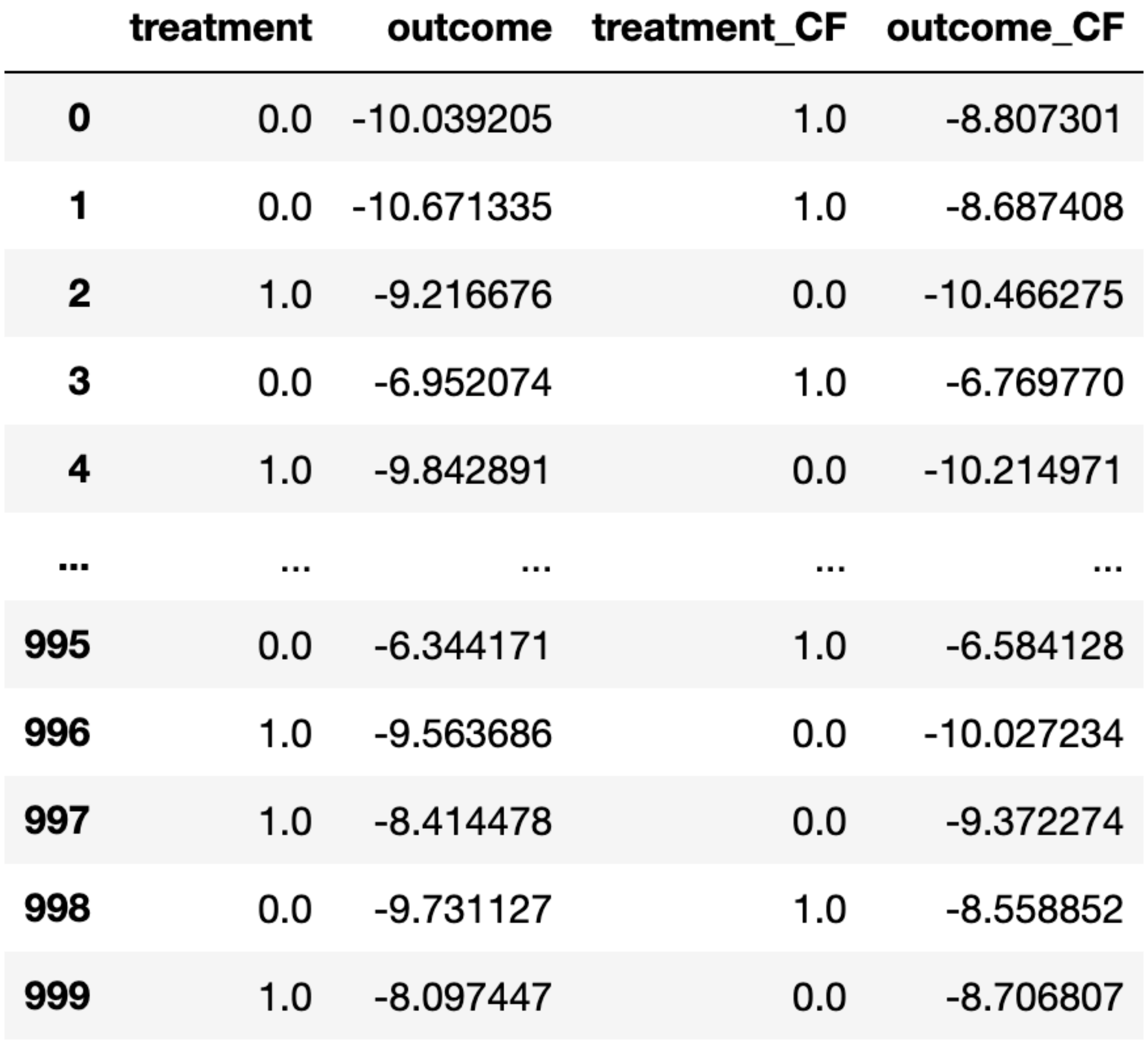
	treatment	outcome	Y_0	Y_1	Y_1-Y_0
0	0.0	-10.039205	-10.039205	?	?
1	0.0	-10.671335	-10.671335	?	?
2	1.0	-9.216676	?	-9.216676	?
3	0.0	-6.952074	-6.952074	?	?
4	1.0	-9.842891	?	-9.842891	?
...	...	...	...	...	...
995	0.0	-6.344171	-6.344171	?	?
996	1.0	-9.563686	?	-9.563686	?
997	1.0	-8.414478	?	-8.414478	?
998	0.0	-9.731127	-9.731127	?	?
999	1.0	-8.097447	?	-8.097447	?

What about the naive observational estimator?

$$\underbrace{\mathbb{E}[Y|T=1]}_{-9.70} - \underbrace{\boxed{\mathbb{E}[Y|T=0]}}_{-8.55}$$

= -1.14

Example (Missing data interpretation)

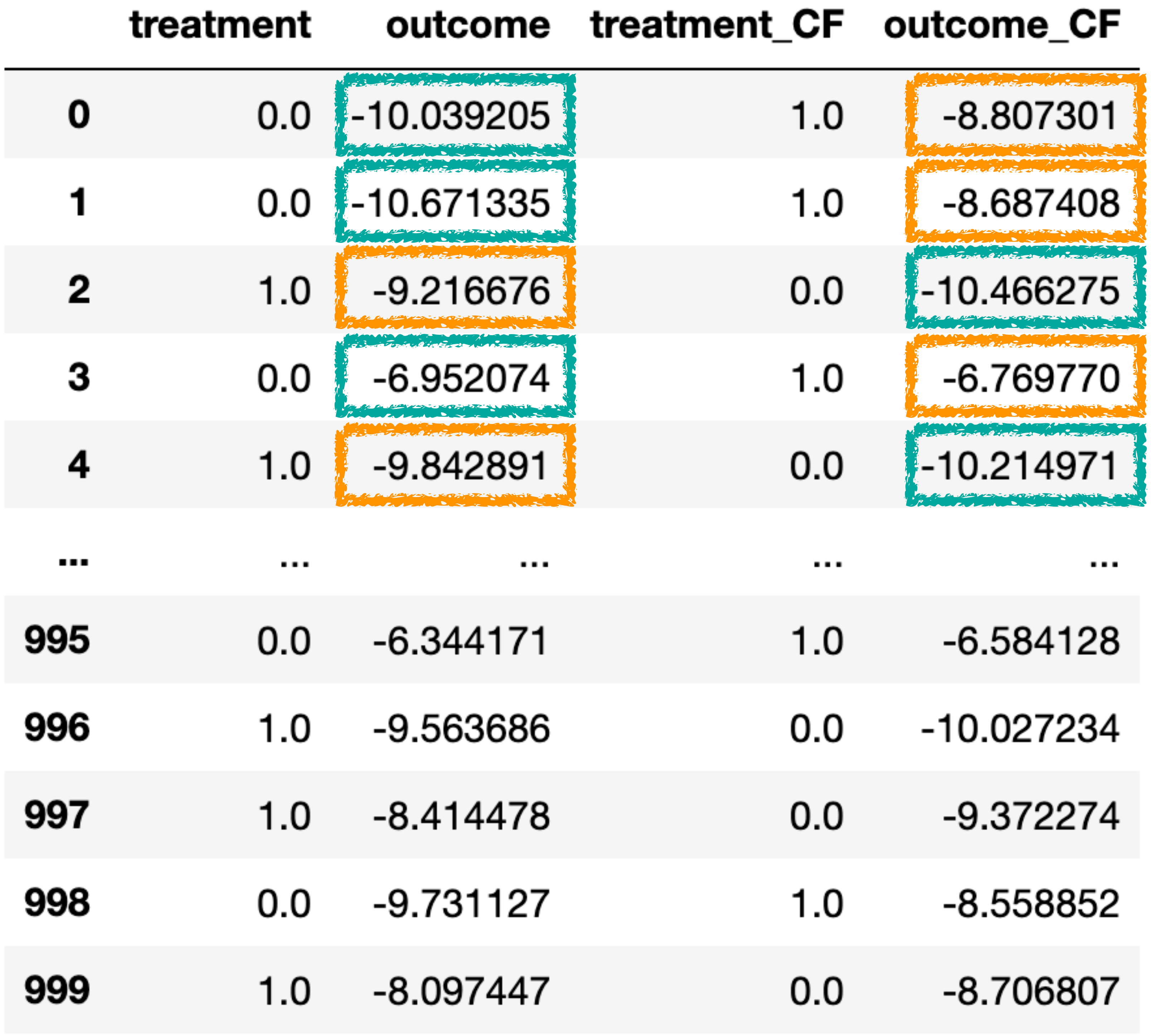


	treatment	outcome	treatment_CF	outcome_CF
0	0.0	-10.039205	1.0	-8.807301
1	0.0	-10.671335	1.0	-8.687408
2	1.0	-9.216676	0.0	-10.466275
3	0.0	-6.952074	1.0	-6.769770
4	1.0	-9.842891	0.0	-10.214971
...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128
996	1.0	-9.563686	0.0	-10.027234
997	1.0	-8.414478	0.0	-9.372274
998	0.0	-9.731127	1.0	-8.558852
999	1.0	-8.097447	0.0	-8.706807

Individual treatment effect:

$$\mathbb{E}[Y_1 - Y_0]$$

Example (Missing data interpretation)



	treatment	outcome	treatment_CF	outcome_CF
0	0.0	-10.039205	1.0	-8.807301
1	0.0	-10.671335	1.0	-8.687408
2	1.0	-9.216676	0.0	-10.466275
3	0.0	-6.952074	1.0	-6.769770
4	1.0	-9.842891	0.0	-10.214971
...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128
996	1.0	-9.563686	0.0	-10.027234
997	1.0	-8.414478	0.0	-9.372274
998	0.0	-9.731127	1.0	-8.558852
999	1.0	-8.097447	0.0	-8.706807

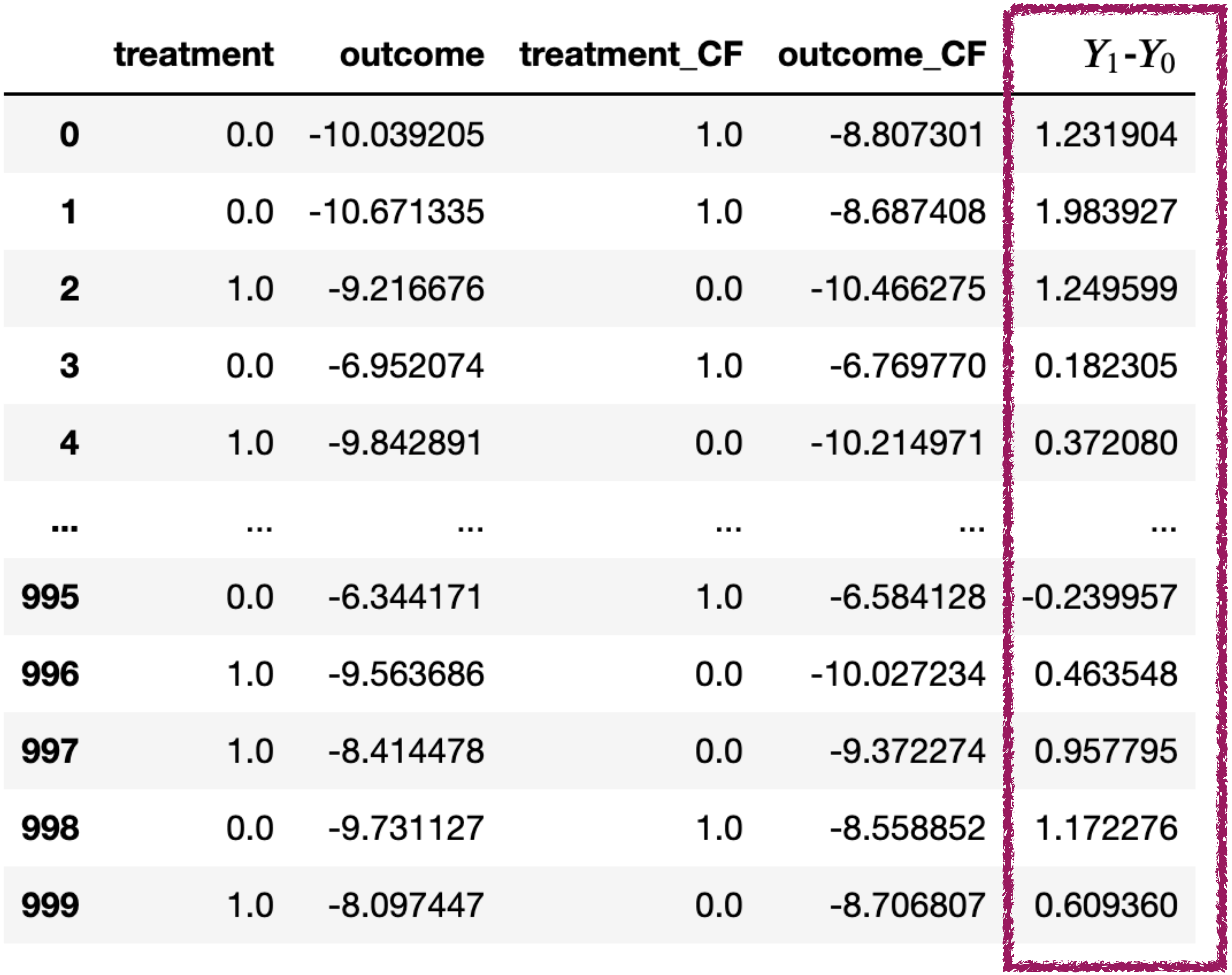
Individual treatment effect:

$$\mathbb{E}[Y_1 - Y_0]$$

Estimated as:

$$\frac{1}{N}\sum_{i=0}^{N}\left(\boxed{y_1^{(i)}} - \boxed{y_0^{(i)}}\right)$$

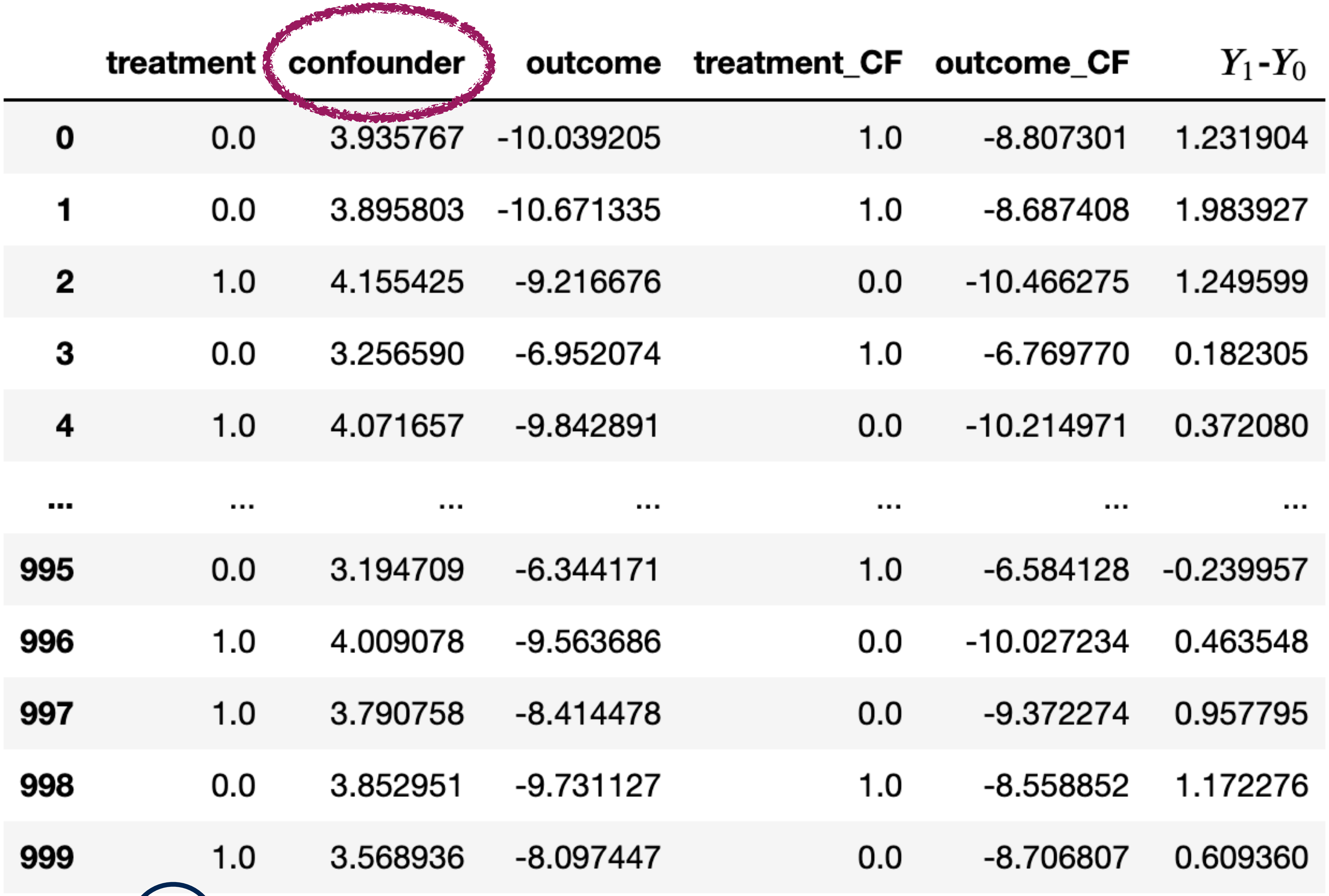
Example (Missing data interpretation)



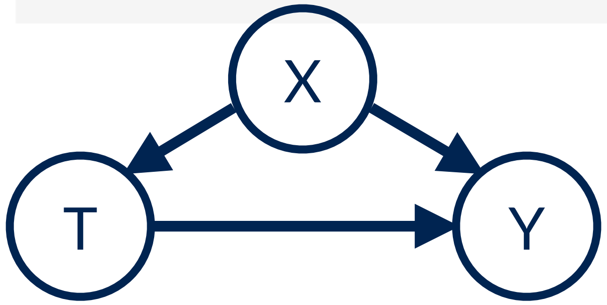
	treatment	outcome	treatment_CF	outcome_CF	Y_1-Y_0
0	0.0	-10.039205	1.0	-8.807301	1.231904
1	0.0	-10.671335	1.0	-8.687408	1.983927
2	1.0	-9.216676	0.0	-10.466275	1.249599
3	0.0	-6.952074	1.0	-6.769770	0.182305
4	1.0	-9.842891	0.0	-10.214971	0.372080
...	...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128	-0.239957
996	1.0	-9.563686	0.0	-10.027234	0.463548
997	1.0	-8.414478	0.0	-9.372274	0.957795
998	0.0	-9.731127	1.0	-8.558852	1.172276
999	1.0	-8.097447	0.0	-8.706807	0.609360

✓ 1.00 $\neq$ -1.14 ⚠

Example (Missing data interpretation)



	treatment	confounder	outcome	treatment_CF	outcome_CF	Y_1-Y_0
0	0.0	3.935767	-10.039205	1.0	-8.807301	1.231904
1	0.0	3.895803	-10.671335	1.0	-8.687408	1.983927
2	1.0	4.155425	-9.216676	0.0	-10.466275	1.249599
3	0.0	3.256590	-6.952074	1.0	-6.769770	0.182305
4	1.0	4.071657	-9.842891	0.0	-10.214971	0.372080
...	...	...	...	...	...	...
995	0.0	3.194709	-6.344171	1.0	-6.584128	-0.239957
996	1.0	4.009078	-9.563686	0.0	-10.027234	0.463548
997	1.0	3.790758	-8.414478	0.0	-9.372274	0.957795
998	0.0	3.852951	-9.731127	1.0	-8.558852	1.172276
999	1.0	3.568936	-8.097447	0.0	-8.706807	0.609360



Potential Outcomes: Assumptions

- **SUTVA:** Stable Unit Treatment Value Assumption
  - **Consistency:** Well-defined treatment (no different versions) potential outcome is independent of how the treatment is assigned
  - **No interference**: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

Potential Outcomes: Assumptions

- **SUTVA:** Stable Unit Treatment Value Assumption
  - **Consistency:** Well-defined treatment (no different versions) potential outcome is independent of how the treatment is assigned
  - **No interference**: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

- **Positivity:** Every individual has a non-zero chance of receiving the treatment/control:

$$p(t = 1|x) \in (0, 1) \text{ if } P(x) > 0$$

- **Unconfoundedness:** Treatment assignment is random, given confounding features X

Unconfoundedness

- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

- Given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)

Unconfoundedness

- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

- Given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person

Unconfoundedness

- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

- Given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person
- There may be difference in sample size between case and control:

not necessarily = $p(t = 0|x)$

Unconfoundedness

- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

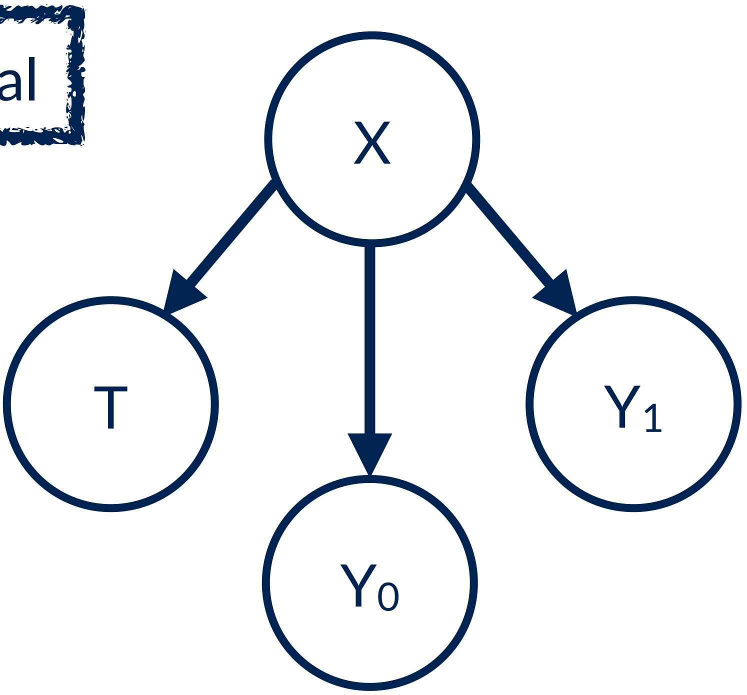
- Given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person
- There may be difference in sample size between case and control: not necessarily = $p(t = 0|x)$
- However, if we do not restrict to the old group, there is a clear preference: older individuals are more likely to receive the drug
- **No unobserved confounders** (see later: unverifiable in observational data)

Unconfoundedness: A graphical representation

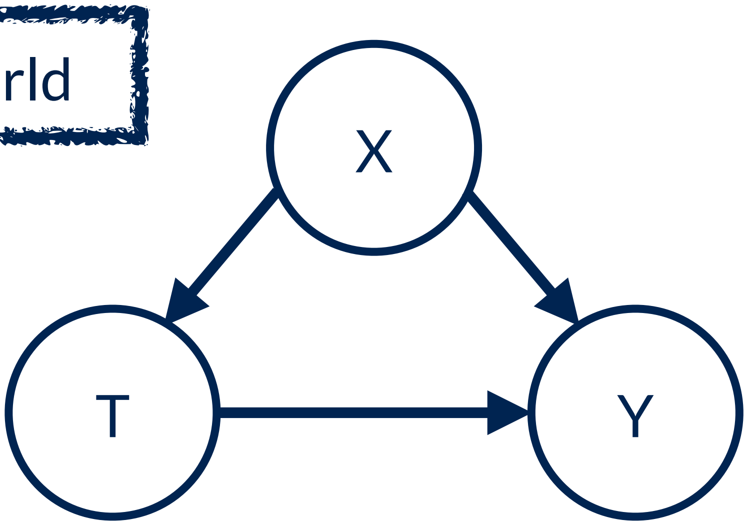
- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

Hypothetical



Real world



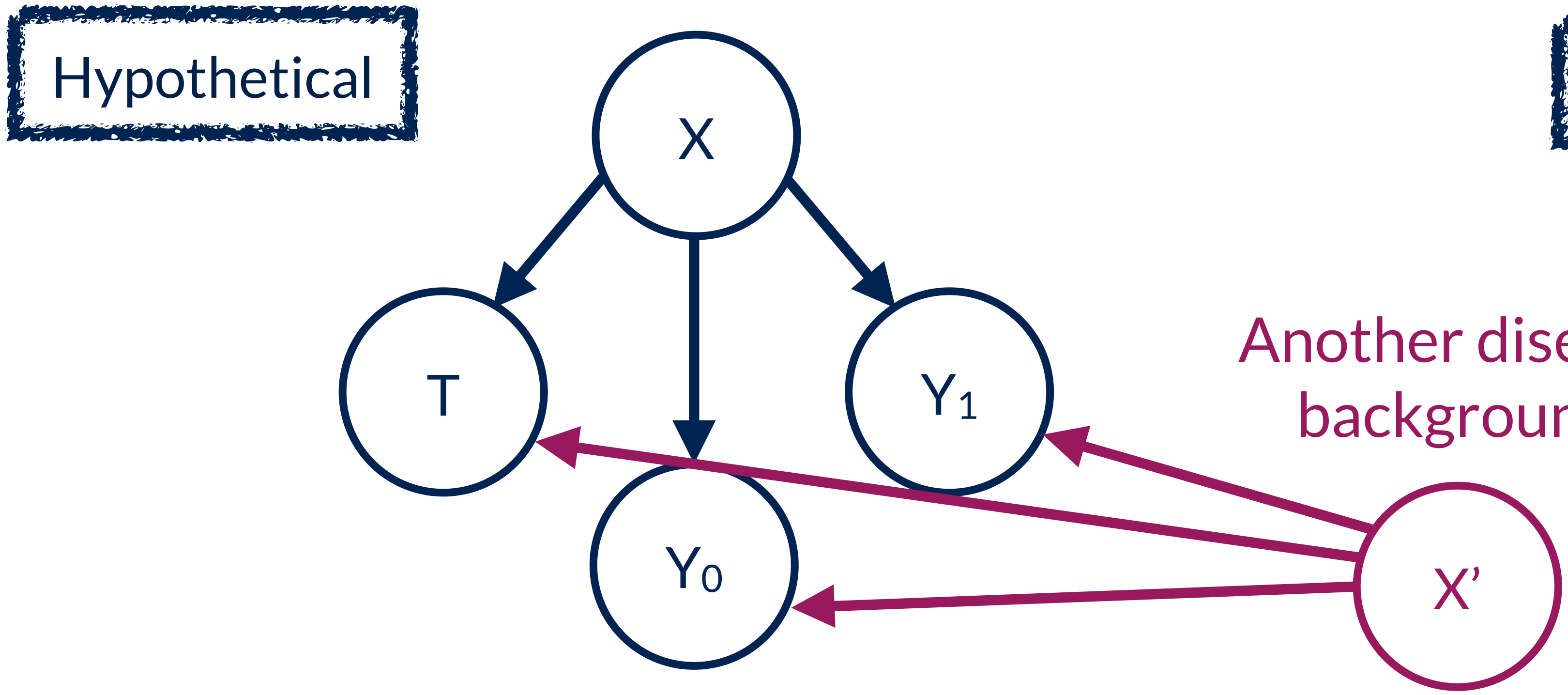
If everyone receive the treatment: Y_1
If everyone is prevented from receiving the treatment: Y_0

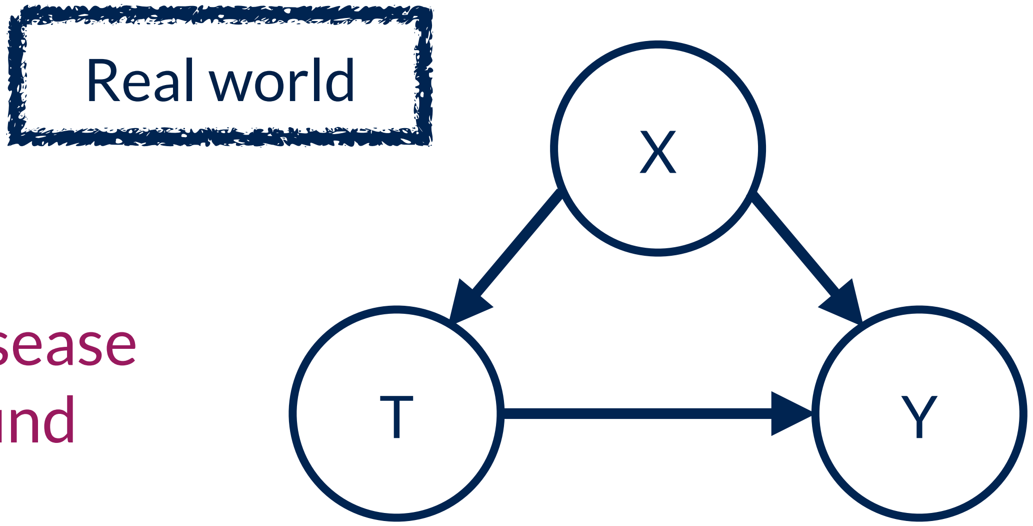
Then the hypothetical outcomes are entirely determined by the set of features X of the individuals.

Unconfoundedness: A graphical representation

- **Unconfoundedness:** Treatment assignment is random, given X:

$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$





$$y_1^{(i)}, y_0^{(i)} \not\!\perp\!\!\!\perp t^{(i)} \mid x$$

If everyone receive the treatment: Y_1
If everyone is prevented from receiving the treatment: Y_0

Then the hypothetical outcomes are entirely determined by the set of features X of the individuals.

Positivity

For existing values of covariates in the population, i.e., $P(X = x) > 0$ (binary T)

$$\boxed{0 < P(T = 1|X = x) < 1}$$

Intuitively: If everyone was given the treatment, i.e., there is no control group, we have no idea if/how the outcomes observed are due to the treatment itself (because we have no background to compare it to!)

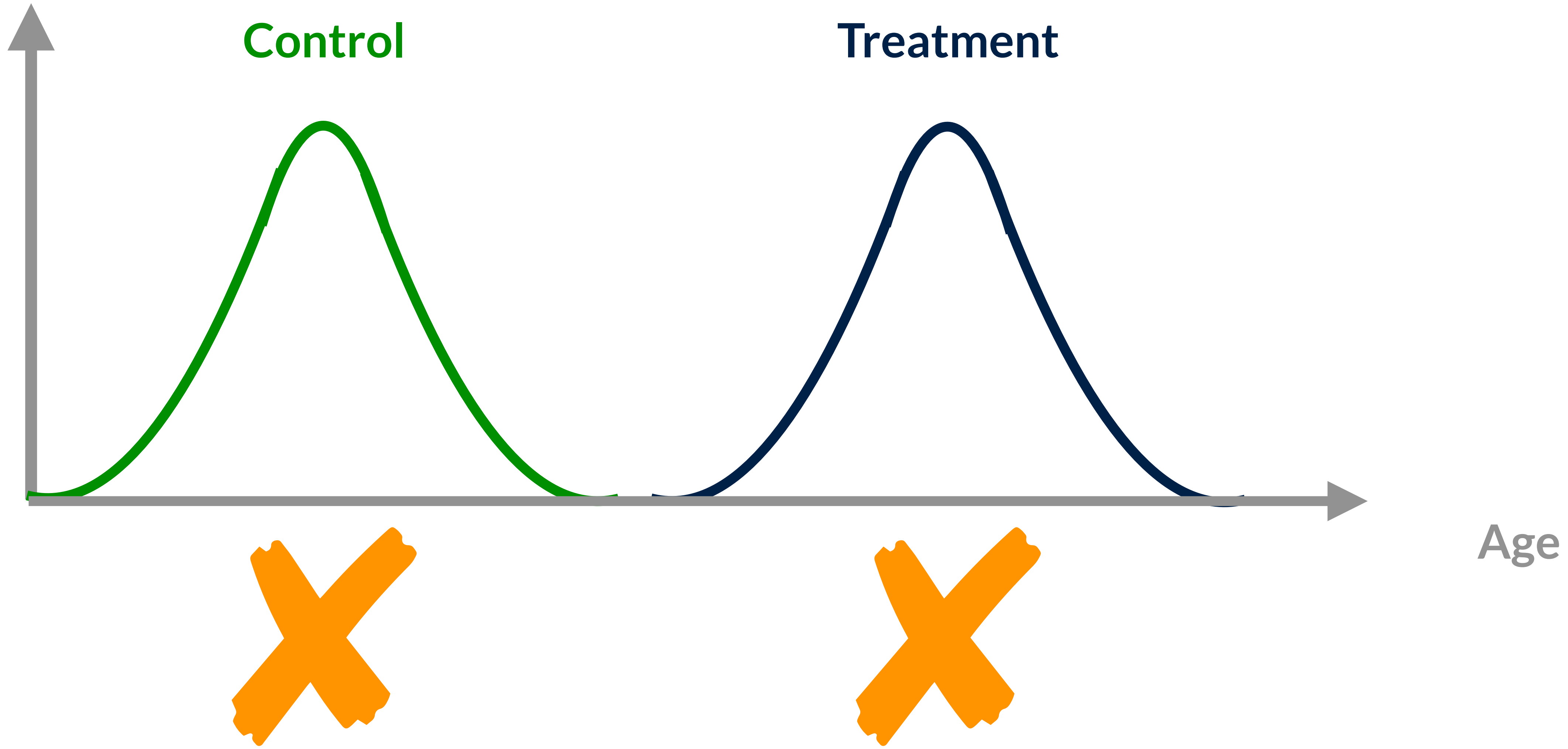
Similarly, when everyone is in the control group: Then we will not have tested the treatment.

Tutorial question: See why this condition is essential (**mathematically**)

Positivity (common support/overlap)

Control: T = 0

Treatment T=1



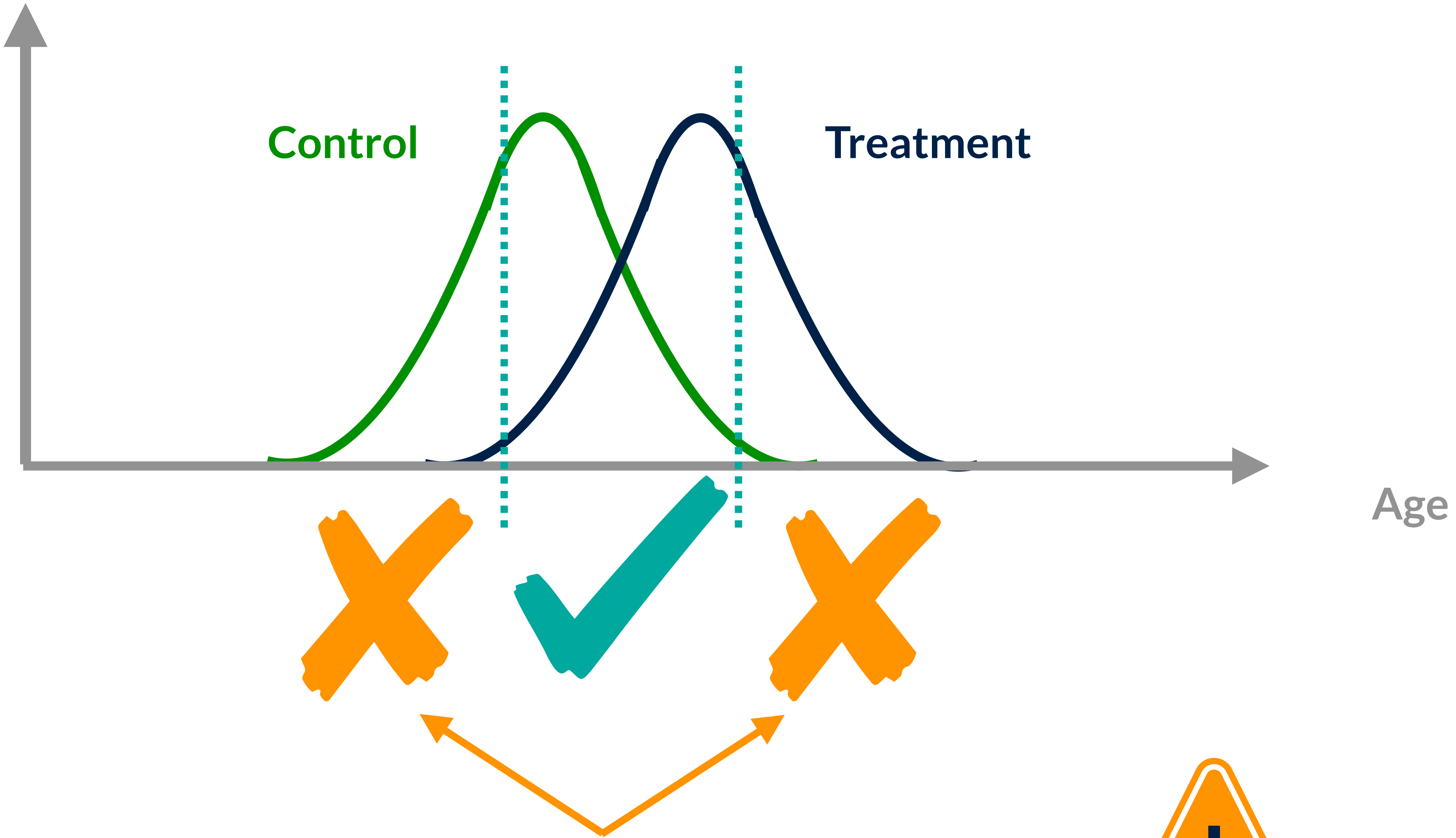
No overlap

Complete violation of positivity

Positivity (common support/overlap)

Control: T = 0

Treatment T=1

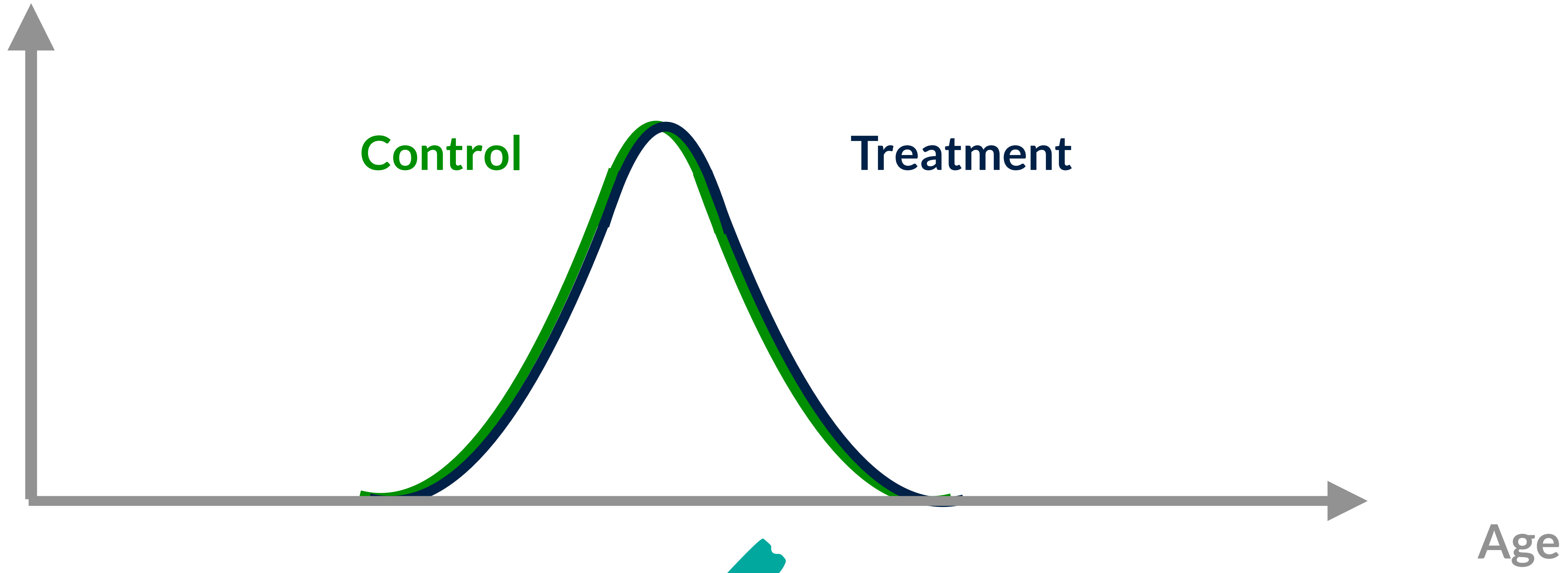


Some violation of positivity

Positivity (common support/overlap)

Control: T = 0

Treatment T=1



✓

Complete overlap: No positivity violation

Positivity vs unconfoundedness

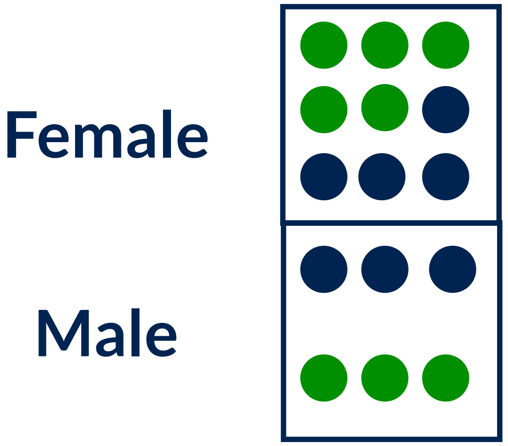
Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

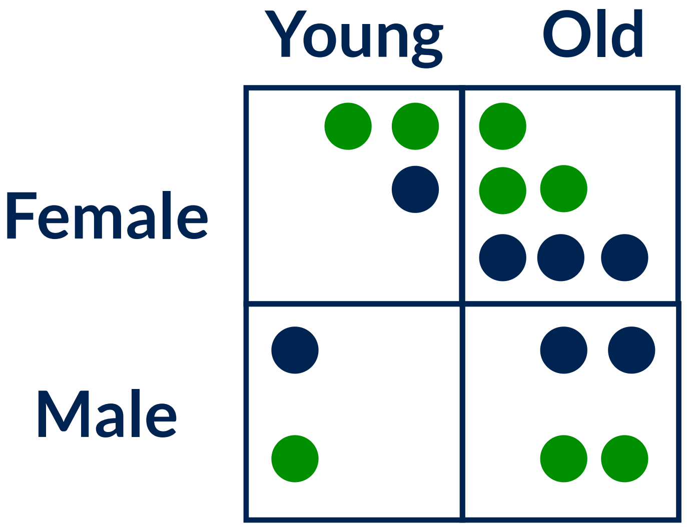
Positivity vs unconfoundedness

Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

But the more we condition on, the harder it is to satisfy positivity

Example:





Easy to check for binary/categorical variable X:

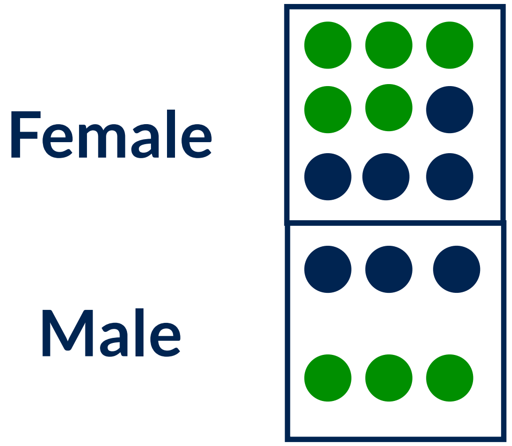
$$0 < P(T = 1|X = x) < 1$$

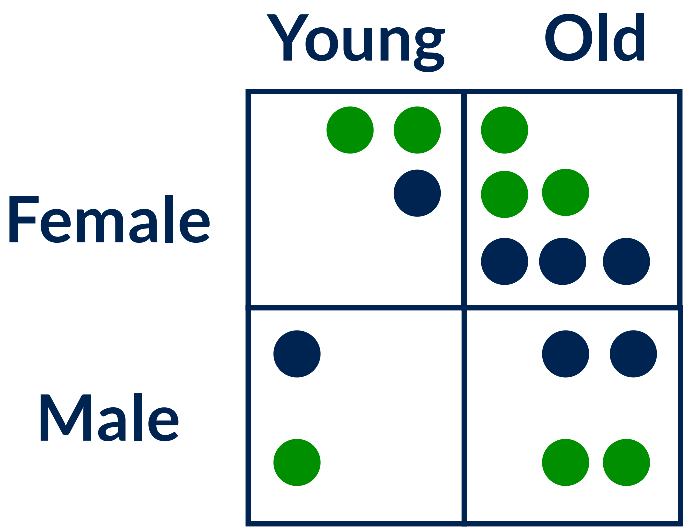
Positivity vs unconfoundedness

Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

But the more we condition on, the harder it is to satisfy positivity

Example:



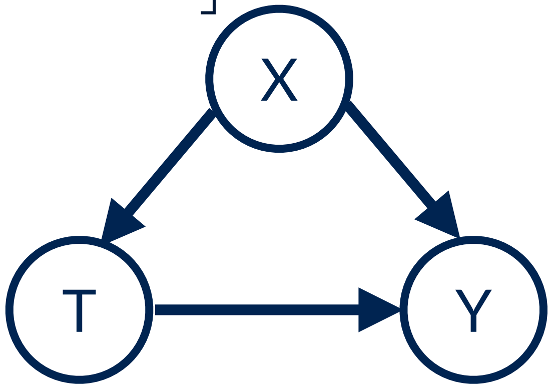


Tutorial question: Discuss the problem of no support, extrapolation and model-misspecification

Regression Adjustment

- X is a sufficient set of confounders if conditioning on X, there would be no confounding bias

- For individual (i) there is only one **observed** outcome: $y_{t_i}^{(i)}$
- Would like to estimate (infer) **counterfactual**: $\hat{y}_{1-t_i}^{(i)} = \hat{\mathbb{E}}\left[y^{(i)} | 1 - t_i, x^{(i)}\right]$



- Using a design matrix, fit: $Y = \beta_X X + \beta_T T + \epsilon$

$$T = \begin{array}{cc} \textbf{Ctrl} & \textbf{Drug} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \qquad X = \begin{array}{cc} \textbf{Young} & \textbf{Old} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ .. & .. \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \Longrightarrow \quad \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ .. \\ y^{(N-1)} \\ y^{(N)} \end{pmatrix} = \begin{pmatrix} \beta_{t=0} + \beta_{x=\text{young}} \\ \beta_{t=0} + \beta_{x=\text{old}} \\ .. \\ \beta_{t=1} + \beta_{x=\text{young}} \\ \beta_{t=1} + \beta_{x=\text{old}} \end{pmatrix}$$

- Assumptions: Overlap and additivity

$$\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N} \sum_{i=0}^{N} \left(y_1^{(i)} - y_0^{(i)} \right)$$

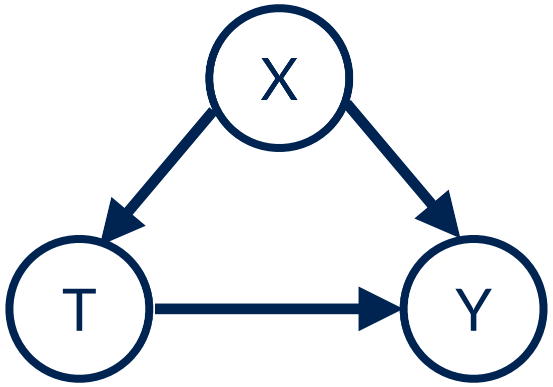
Adjustment formula (will be revisited later)

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|X] =& \mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X] \\ =& \mathbb{E}[Y_1|T=1,X] - \mathbb{E}[Y_0|T=0,X] \\ =& \mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\end{aligned}$$

By Unconfoundedness: $Y_1, Y_0 \perp\!\!\!\perp T \mid X$

By construction: $Y = TY_1 + (1-T)Y_0$

Also need positivity

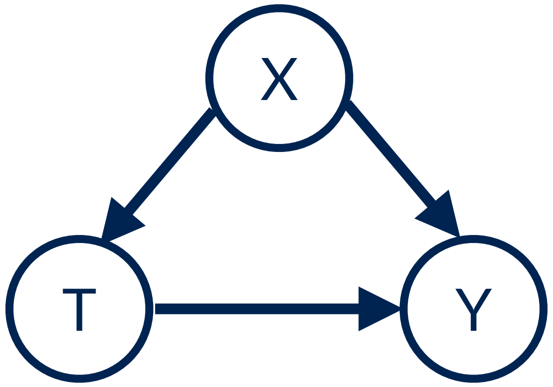


Adjustment formula (will be revisited later)

$$\begin{aligned}
\mathbb{E}[Y_1 - Y_0|X] =& \mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X] \\
=& \mathbb{E}[Y_1|T=1,X] - \mathbb{E}[Y_0|T=0,X] && \text{By Unconfoundedness:} \quad Y_1, Y_0 \perp\!\!\!\perp T \mid X \\
=& \mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X] && \text{By construction:} \quad Y = TY_1 + (1-T)Y_0
\end{aligned}$$

Also need positivity

$$\begin{aligned}
\mathbb{E}[Y_1 - Y_0] =& \mathbb{E}_X\Big[\mathbb{E}[Y_1 - Y_0|X]\Big] \\
\boxed{\text{ATE}} \qquad =& \mathbb{E}_X\Big[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\Big] && \boxed{\text{The adjustment formula}}
\end{aligned}$$



Adjustment formula (will be revisited later)

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|X] =& \mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X] \\ =& \mathbb{E}[Y_1|T=1, X] - \mathbb{E}[Y_0|T=0, X] \\ =& \mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]\end{aligned}$$

By Unconfoundedness: $Y_1, Y_0 \perp\!\!\!\perp T \mid X$

By construction: $Y = TY_1 + (1-T)Y_0$

Also need positivity

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0] =& \mathbb{E}_X\Big[\mathbb{E}[Y_1 - Y_0|X]\Big] \\ =& \underbrace{\mathbb{E}_X\Big[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]\Big]}\end{aligned}$$

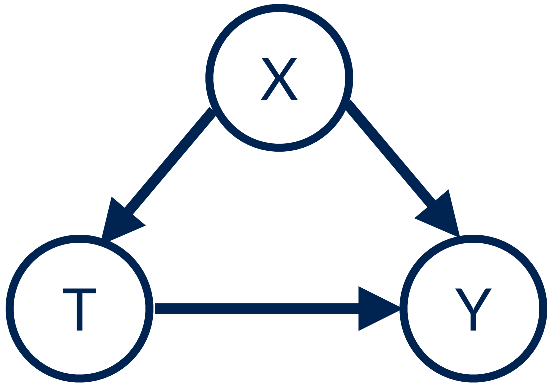
The adjustment formula

Hypothetical world

Real world

i.e., can be estimated from observational data

Causal identifiability



Regression Adjustment: Another perspective

Fit a model for $Q(T, X) = \mathbb{E}[Y|T, X]$

(last time we substituted T=1 and T=0 into individual treatment effect = $Q(1, x^{(i)}) - Q(0, x^{(i)})$, then took average over all individuals i, via linear regression). Under the linearity assumption:

$$\mathbb{E}[Y|T, X] = \alpha_0 + \beta_x X + \beta_t T + \epsilon \ , \ \mathbb{E}[\epsilon] = 0$$

Regression Adjustment: Another perspective

Fit a model for $Q(T, X) = \mathbb{E}[Y|T, X]$

(last time we substituted T=1 and T=0 into individual treatment effect $= Q(1, x^{(i)}) - Q(0, x^{(i)})$, then took average over all individuals i, via linear regression). Under the linearity assumption:

$$\mathbb{E}[Y|T, X] = \alpha_0 + \beta_x X + \beta_t T + \epsilon \ , \ \mathbb{E}[\epsilon] = 0$$

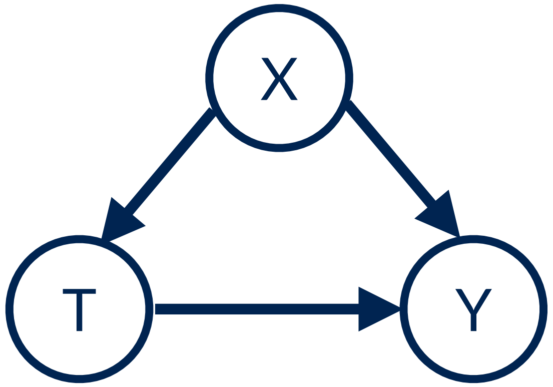
$$\begin{aligned} ATE &= \mathbb{E}_X\Big[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]\Big] \\ &= \Big(\cancel{\alpha_0 + \beta_x \mathbb{E}[X]} + \beta_t\Big) - \Big(\cancel{\alpha_0 + \beta_x \mathbb{E}[X]}\Big) \\ &= \beta_t \end{aligned}$$

Important remarks about the previous form:

1) Depends on the structure of the causal graph of interest

2) Data need not be linear

   model-misspecification -> statistical bias



Important remarks about the previous form:

2) Data need not be linear, example:

Say we fitted $\mathbb{E}[Y|T,X] = \alpha_0 + \beta_x X + \beta_t T + \epsilon\ ,\ \mathbb{E}[\epsilon] = 0$
And obtained $\beta_t$for the causal effect,

BUT, in reality the true data generating distribution is e.g.

$$\mathbb{E}[Y|T,X] = \alpha_0 + \beta_x X + \beta_t T + \gamma X.T + \epsilon\ ,\ \mathbb{E}[\epsilon] = 0$$

Or e.g. non-linear:

$$\mathbb{E}[Y|T,X] = e^{\alpha_0 + \beta_x X + \beta_t T + \gamma X.T}$$

Important remarks about the previous form:

2) Data need not be linear, example:

Say we fitted $\mathbb{E}[Y|T,X] = \alpha_0 + \beta_x X + \beta_t T + \epsilon \ , \ \mathbb{E}[\epsilon] = 0$
And obtained $\beta_t$for the causal effect,

BUT, in reality the true data generating distribution is e.g.

$$\mathbb{E}[Y|T,X] = \alpha_0 + \beta_x X + \beta_t T + \gamma X.T + \epsilon \ , \ \mathbb{E}[\epsilon] = 0$$

Or e.g. non-linear:

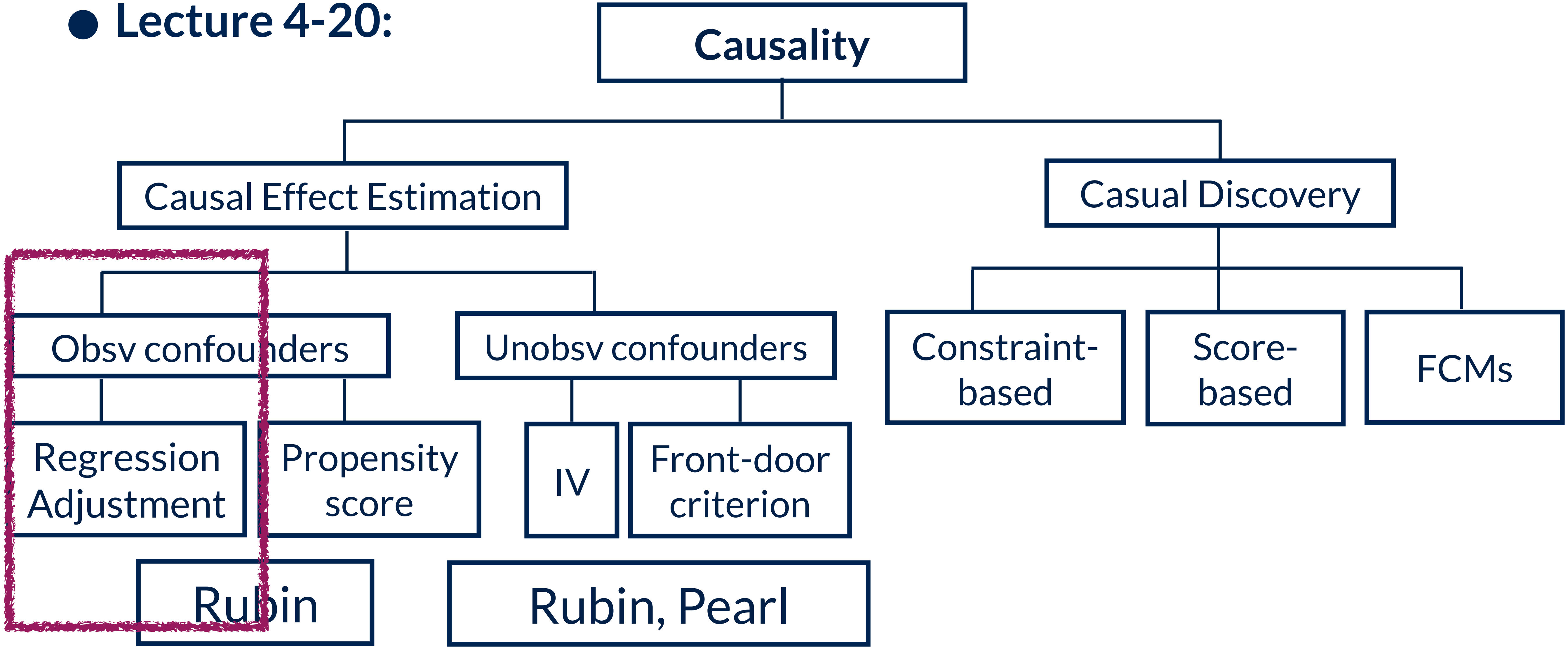
$$\mathbb{E}[Y|T,X] = e^{\alpha_0 + \beta_x X + \beta_t T + \gamma X.T}$$

Then $ATE = \mathbb{E}_X \Big[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X] \Big]$
is **not** simply β_t !!

Valid causal inference requires correctly-specified models and mathematical guarantees!

Overview of the course

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**



THE UNIVERSITY *of* EDINBURGH

Methods for Causal Inference
Lecture 4: Rubin's potential outcomes framework

Ava Khamseh

School of Informatics

2024-2025