Czy witaminy leczą raka?
Oszustwa w modelowaniu przyczynowości

Krzysztof Rudaś [1,2]

[1] Politechnika Warszawska

[2] Instytut Podstaw Informatyki PAN

LX Szkoła Matematyki Poglądowej

Wola Ducka 26.08.2019

Początki modelowania przyczynowości

Czym jest modelowanie przyczynowości?

- W życiu codziennym spotykamy się z formułowaniem przyczynowości.

Czym jest modelowanie przyczynowości?

- W życiu codziennym spotykamy się z formułowaniem przyczynowości.
  - "Przestała mnie boleć głowa bo wziąłem aspirynę."
  - "Miałbym mniej płatną pracę, gdybym nie poszedł na studia."
  - "Mam krótkie włosy ponieważ jestem mężczyzną."

Czym jest modelowanie przyczynowości?

- W życiu codziennym spotykamy się z formułowaniem przyczynowości.
  - "Przestała mnie boleć głowa bo wziąłem aspirynę."
  - "Miałbym mniej płatną pracę, gdybym nie poszedł na studia."
  - "Mam krótkie włosy ponieważ jestem mężczyzną."
- Czy wszystkie te sformułowania da się opisać statystycznie?

Czym jest modelowanie przyczynowości?

- W życiu codziennym spotykamy się z formułowaniem przyczynowości.
  - "Przestała mnie boleć głowa bo wziąłem aspirynę."
  - "Miałbym mniej płatną pracę, gdybym nie poszedł na studia."
  - "Mam krótkie włosy ponieważ jestem mężczyzną."
- Czy wszystkie te sformułowania da się opisać statystycznie?
- Aby na to odpowiedzieć musimy wprowadzić kilka pojęć.

Działanie

- Akcja, którą możemy podjąć lub nie.

Działanie

- Akcja, którą możemy podjąć lub nie.
  - Wzięcie aspiryny,
  - Pójście na studia,
  - Stanie się mężczyzną,

Działanie

- Akcja, którą możemy podjąć lub nie.
  - Wzięcie aspiryny,
  - Pójście na studia,
  - Stanie się mężczyzną,
- Działanie powinno być zdefiniowane jak najdokładniej.

Potencjalny stan kontrfaktyczny

- Alternatywny stan gdybyśmy nie podjęli działania.

Potencjalny stan kontrfaktyczny

- Alternatywny stan gdybyśmy nie podjęli działania.
  - Ból głowy (lub jego brak) gdybym nie wziął aspiryny,
  - Zarobki, gdybym nie poszedł na studia,
  - Długość włosów gdybym był kobietą.

Potencjalny stan kontrfaktyczny

- Alternatywny stan gdybyśmy nie podjęli działania.
  - Ból głowy (lub jego brak) gdybym nie wziął aspiryny,
  - Zarobki, gdybym nie poszedł na studia,
  - Długość włosów gdybym był kobietą.
- Przyczynowość możemy analizować statystycznie tylko wtedy gdy stan kontrfaktyczny jest:

Potencjalny stan kontrfaktyczny

- Alternatywny stan gdybyśmy nie podjęli działania.
  - Ból głowy (lub jego brak) gdybym nie wziął aspiryny,
  - Zarobki, gdybym nie poszedł na studia,
  - Długość włosów gdybym był kobietą.
- Przyczynowość możemy analizować statystycznie tylko wtedy gdy stan kontrfaktyczny jest:
  - Jasno określony (pkt. 2),

Potencjalny stan kontrfaktyczny

- Alternatywny stan gdybyśmy nie podjęli działania.
  - Ból głowy (lub jego brak) gdybym nie wziął aspiryny,
  - Zarobki, gdybym nie poszedł na studia,
  - Długość włosów gdybym był kobietą.
- Przyczynowość możemy analizować statystycznie tylko wtedy gdy stan kontrfaktyczny jest:
  - Jasno określony (pkt. 2),
  - Możliwy do uzyskania (pkt. 3).

Sformułowanie problemu

- Przyczynowość można badać na wiele różnych sposobów.

Sformułowanie problemu

- Przyczynowość można badać na wiele różnych sposobów.
- Średni efekt akcji.

Sformułowanie problemu

- Przyczynowość można badać na wiele różnych sposobów.
- Średni efekt akcji.
  - Porównujemy średni efekt na populacji poddanej działaniu i nie poddanej.

Sformułowanie problemu

- Przyczynowość można badać na wiele różnych sposobów.
- Średni efekt akcji.
  - Porównujemy średni efekt na populacji poddanej działaniu i nie poddanej.
  - Możliwe warstwowanie - można przeprowadzić oddzielne badania dla kobiet i mężczyzn i wnioskować czy płeć ma wpływ na efektywność działania.

Sformułowanie problemu

- Przyczynowość można badać na wiele różnych sposobów.
- Średni efekt akcji.
  - Porównujemy średni efekt na populacji poddanej działaniu i nie poddanej.
  - Możliwe warstwowanie - można przeprowadzić oddzielne badania dla kobiet i mężczyzn i wnioskować czy płeć ma wpływ na efektywność działania.
- Wybór osób, które powinny zostać poddane działaniu.

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,
- Naszych pacjentów możemy podzielić na trzy grupy

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,
- Naszych pacjentów możemy podzielić na trzy grupy
  1. pacjenci, którym spadło ciśnienie **na skutek** działania leku (**pozytywny efekt**)

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,
- Naszych pacjentów możemy podzielić na trzy grupy
  1. pacjenci, którym spadło ciśnienie **na skutek** działania leku (**pozytywny efekt**)
  2. pacjenci, którym ciśnienie się nie zmieniło **mimo** działania leku (**niepotrzebne działanie**)

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,
- Naszych pacjentów możemy podzielić na trzy grupy
  1. pacjenci, którym spadło ciśnienie **na skutek** działania leku (**pozytywny efekt**)
  2. pacjenci, którym ciśnienie się nie zmieniło **mimo** działania leku (**niepotrzebne działanie**)
  3. pacjenci, którym ciśnienie wzrosło **na skutek** działania leku (**negatywny efekt**)

Wybór osób - przykład

- Załóżmy że testujemy nowy lek na nadciśnienie (każdy pacjent dostaje tą samą dawkę),
- Chcemy ocenić którym pacjentom jest w stanie pomóc,
- Naszych pacjentów możemy podzielić na trzy grupy
  1. pacjenci, którym spadło ciśnienie **na skutek** działania leku (**pozytywny efekt**)
  2. pacjenci, którym ciśnienie się nie zmieniło **mimo** działania leku (**niepotrzebne działanie**)
  3. pacjenci, którym ciśnienie wzrosło **na skutek** działania leku (**negatywny efekt**)
- Cel: Znalezienie osób, które powinny dostać lek.

Powszechnie stosowane metody

- Lek powinien być przeznaczony dla osób o największej różnicy ciśnienia po zażyciu leku.

Różnica między ciśnieniem po i przed eksperymentem:	
	gdy pacjent otrzymał lek
Pacjent 1	+10
Pacjent 2	-20
Pacjent 3	-20
Pacjent 4	0

Powszechnie stosowane metody

- Lek powinien być przeznaczony dla osób o największej różnicy ciśnienia po zażyciu leku.

Różnica między ciśnieniem po i przed eksperymentem:	
	gdy pacjent otrzymał lek
Pacjent 1	+10
Pacjent 2	-20
Pacjent 3	-20
Pacjent 4	0

Powszechnie stosowane metody

- Lek powinien być przeznaczony dla osób o największej różnicy ciśnienia po zażyciu leku.

Różnica między ciśnieniem po i przed eksperymentem:		
	gdy pacjent nie otrzymał leku	gdy pacjent otrzymał lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Lek jest aplikowany dla osób o największej różnicy ciśnień po i przed podaniem leku, a nie dla osób, dla których ciśnienie spadło znacząco na skutek działania leku.

Powszechnie stosowane metody

- Lek powinien być przeznaczony dla osób o największej różnicy ciśnienia po zażyciu leku.

Różnica między ciśnieniem po i przed eksperymentem:		
	gdy pacjent nie otrzymał leku	gdy pacjent otrzymał lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Lek jest aplikowany dla osób o największej różnicy ciśnień po i przed podaniem leku, a nie dla osób, dla których ciśnienie spadło znacząco na skutek działania leku.

Podstawowy problem modelowania przyczynowości

Chcemy prognozować:

$$\mathrm{E}^{Target}(Y \mid X_1, \ldots, X_p) - \mathrm{E}^{Not\ target}(Y \mid X_1, \ldots, X_p)$$

Problem: Nigdy nie mamy tych dwóch informacji w tym samym czasie.

Podstawowy problem modelowania przyczynowości

Chcemy prognozować:

$$\mathrm{E}^{Target}(Y \mid X_1, \ldots, X_p) - \mathrm{E}^{Not\ target}(Y \mid X_1, \ldots, X_p)$$

Problem: Nigdy nie mamy tych dwóch informacji w tym samym czasie.

Różnica między ciśnieniem po i przed eksperymentem:		
	gdy pacjent nie otrzymał leku	gdy pacjent otrzymał lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

Rozwiązanie problemu

Rozwiązanie:

- Dwa zbiory uczące:
  1. grupa eksperymentalna
     poddana działaniu
  2. grupa kontrolna
     nie poddana działaniu
     używana jako punkt odniesienia

- Skonstruujmy model prognozujący różnicę pomiędzy wartościami w grupie eksperymentalnej i kontrolnej.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.
- Pacjenci przebywają w jednym pomieszczeniu.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.
- Pacjenci przebywają w jednym pomieszczeniu.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.
- Pacjenci przebywają w jednym pomieszczeniu.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.
- Pacjenci przebywają w jednym pomieszczeniu.

Czy to wystarczy?

- Testujemy lek na grupie eksperymentalnej.
- Pacjenci przebywają w jednym pomieszczeniu.

- Pacjenci nie mogą na siebie wpływać.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

	Czy otrzymał lek	Wartość różnicy ciśnień
Pacjent 1	1	+10
Pacjent 2	0	-30
Pacjent 3	0	-20
Pacjent 4	1	0

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

	Czy otrzymał lek	Wartość różnicy ciśnień
Pacjent 1	1	+10
Pacjent 2	0	-30
Pacjent 3	0	-20
Pacjent 4	1	0

- Różnica między średnimi z grup eksperymentalnej i kontrolnej to +30.
- Lek nie pomaga!!!

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie		
	nie otrzymawszy leku	otrzymawszy lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie		
	nie otrzymawszy leku	otrzymawszy lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Różnica między średnimi z grup eksperymentalnej i kontrolnej (przy pełnej informacji) to -10.
- Lek pomaga!!!

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie		
	nie otrzymawszy leku	otrzymawszy lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Różnica między średnimi z grup eksperymentalnej i kontrolnej (przy pełnej informacji) to -10.
- Lek pomaga!!!
- Co jest przyczyną tego paradoksu?

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie		
	nie otrzymawszy leku	otrzymawszy lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Różnica między średnimi z grup eksperymentalnej i kontrolnej (przy pełnej informacji) to -10.
- Lek pomaga!!!
- Co jest przyczyną tego paradoksu?
- Istnieje pewna niewykryta cecha.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie		
	nie otrzymawszy leku	otrzymawszy lek
Pacjent 1	+40	+10
Pacjent 2	-30	-20
Pacjent 3	-20	-20
Pacjent 4	+20	0

- Różnica między średnimi z grup eksperymentalnej i kontrolnej (przy pełnej informacji) to -10.
- Lek pomaga!!!
- Co jest przyczyną tego paradoksu?
- Istnieje pewna niewykryta cecha.
- Wszystkie obserwacje o danej wartości cechy są w grupie kontrolnej, a obserwacje o innej wartości w grupie eksperymentalnej.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie			
	nie otrzymawszy leku	otrzymawszy lek	Płeć
Pacjent 1	+40	+10	M
Pacjent 2	-30	-20	K
Pacjent 3	-20	-20	K
Pacjent 4	+20	0	M

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie			
	nie otrzymawszy leku	otrzymawszy lek	Płeć
Pacjent 1	+40	+10	M
Pacjent 2	-30	-20	K
Pacjent 3	-20	-20	K
Pacjent 4	+20	0	M

- Wpływ leku na nadciśnienie jest zaburzany przez wpływ płci.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie			
	nie otrzymawszy leku	otrzymawszy lek	Płeć
Pacjent 1	+40	+10	M
Pacjent 2	-30	-20	K
Pacjent 3	-20	-20	K
Pacjent 4	+20	0	M

- Wpływ leku na nadciśnienie jest zaburzany przez wpływ płci.
- Sytuacja niebezpieczna, również w przypadku wyboru obserwacji, które powinny być poddane działaniu.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Ciśnienie			
	nie otrzymawszy leku	otrzymawszy lek	Płeć
Pacjent 1	+40	+10	M
Pacjent 2	-30	-20	K
Pacjent 3	-20	-20	K
Pacjent 4	+20	0	M

- Wpływ leku na nadciśnienie jest zaburzany przez wpływ płci.
- Sytuacja niebezpieczna, również w przypadku wyboru obserwacji, które powinny być poddane działaniu.
- Nie mogę powiedzieć czy dany mężczyzna ma dostać lek, skoro nie mam danych, które mówią jak ciśnienie mężczyzn zachowuje się gdy leku nie mają.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

	Czy otrzymał lek	Wartość różnicy ciśnień	Płeć
Pacjent 1	1	+10	M
Pacjent 2	1	-20	K
Pacjent 3	0	-20	K
Pacjent 4	0	+20	M

Czy to wystarczy?

- Rozważmy przykład gdzie badamy średni efekt leku.

	Czy otrzymał lek	Wartość różnicy ciśnień	Płeć
Pacjent 1	1	+10	M
Pacjent 2	1	-20	K
Pacjent 3	0	-20	K
Pacjent 4	0	+20	M

- Różnica między średnimi z grup eksperymentalnej i kontrolnej to -5.
- Uzyskany wynik bardziej odpowiada rzeczywistej sytuacji.

Paradoks Simpsona

- Testujemy lek na nadciśnienie w dwóch grupach (kobiety i mężczyźni),
- Badamy w jakim odsetku przypadków ciśnienie spadło:

Paradoks Simpsona

- Testujemy lek na nadciśnienie w dwóch grupach (kobiety i mężczyźni),
- Badamy w jakim odsetku przypadków ciśnienie spadło:

	Mężczyźni	Kobiety	całość
Lek	60%	31%	33%
Brak leku	44%	20%	42%

Paradoks Simpsona

- Testujemy lek na nadciśnienie w dwóch grupach (kobiety i mężczyźni),
- Badamy w jakim odsetku przypadków ciśnienie spadło:

	Mężczyźni	Kobiety	całość
Lek	60%	31%	33%
Brak leku	44%	20%	42%

- Lek pomaga mężczyznom.
- Lek pomaga kobietom.
- Brak leku lepszy dla całości!!!

Paradoks Simpsona

- Sprawdźmy ile kobiet i mężczyzn otrzymało, bądź nie otrzymało leku.

Paradoks Simpsona

- Sprawdźmy ile kobiet i mężczyzn otrzymało, bądź nie otrzymało leku.

	Mężczyźni	Kobiety	całość
Lek	$\frac{6}{10}$	$\frac{31}{100}$	$\frac{37}{110}$
Brak leku	$\frac{44}{100}$	$\frac{2}{10}$	$\frac{46}{110}$

Paradoks Simpsona

- Sprawdźmy ile kobiet i mężczyzn otrzymało, bądź nie otrzymało leku.

	Mężczyźni	Kobiety	całość
Lek	$\frac{6}{10}$	$\frac{31}{100}$	$\frac{37}{110}$
Brak leku	$\frac{44}{100}$	$\frac{2}{10}$	$\frac{46}{110}$

- inny efekt dla całości spowodowany jest nierównomiernym rozłożeniem kobiet i mężczyzn w grupie biorącej i niebiorącej leku.

Paradoks Simpsona

- Sprawdźmy ile kobiet i mężczyzn otrzymało, bądź nie otrzymało leku.

	Mężczyźni	Kobiety	całość
Lek	$\frac{6}{10}$	$\frac{31}{100}$	$\frac{37}{110}$
Brak leku	$\frac{44}{100}$	$\frac{2}{10}$	$\frac{46}{110}$

- inny efekt dla całości spowodowany jest nierównomiernym rozłożeniem kobiet i mężczyzn w grupie biorącej i niebiorącej leku.

	Mężczyźni	Kobiety	całość
Lek	$\frac{33}{55}$	$\frac{17}{55}$	$\frac{50}{110}$
Brak leku	$\frac{24}{55}$	$\frac{11}{55}$	$\frac{35}{110}$

Randomizacja

- Potrzebny jest mechanizm, który zapewni dobrze przypisanie obserwacji do grup.

Randomizacja

- Potrzebny jest mechanizm, który zapewni dobrze przypisanie obserwacji do grup.
- Rozwiązanie - randomizacja.

Randomizacja

- Potrzebny jest mechanizm, który zapewni dobrze przypisanie obserwacji do grup.
- Rozwiązanie - randomizacja.
- Musimy zastosować procedurę dwustopniową:

Randomizacja

- Potrzebny jest mechanizm, który zapewni dobrze przypisanie obserwacji do grup.
- Rozwiązanie - randomizacja.
- Musimy zastosować procedurę dwustopniową:
  - I stopień - losowy dobór badanej próby z populacji.

Randomizacja

- Potrzebny jest mechanizm, który zapewni dobrze przypisanie obserwacji do grup.
- Rozwiązanie - randomizacja.
- Musimy zastosować procedurę dwustopniową:
  - I stopień - losowy dobór badanej próby z populacji.
  - II stopień - losowy przydział do grupy kontrolnej i eksperymentalnej.

Randomizacja

- Co zyskujemy?

Randomizacja

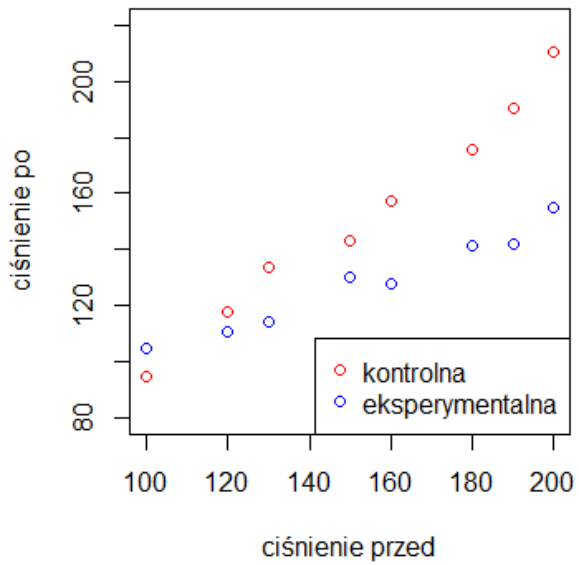
- Co zyskujemy?
  - odzwierciedlenie charakterystyki populacji (np. pacjenci leczący się na nadciśnienie) w badanej próbie (np. 1000 osób)
  - zminimalizowanie wpływu zmiennych na wynik w całej grupie.

Jak modelować?

- Przykład: porównajmy wartość ciśnienia przed działaniem i po działaniu.

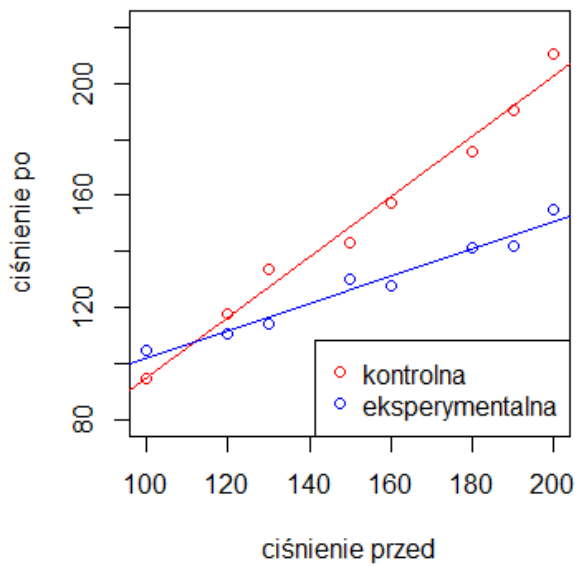
Jak modelować?

- Przykład: porównajmy wartość ciśnienia przed działaniem i po działaniu.



Jak modelować?

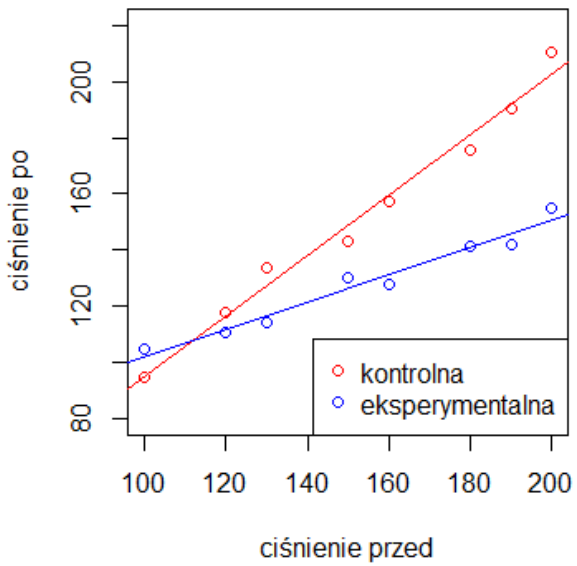
- Przykład: porównajmy wartość ciśnienia przed działaniem i po działaniu.



$y = 1.08x - 12.39$, $y = 0.48x + 53.99$

Jak modelować?

- Przykład: porównajmy wartość ciśnienia przed działaniem i po działaniu.



$y = 1.08x - 12.39$, $y = 0.48x + 53.99$

- Policzmy różnicę na współczynnikach.

$$y = -0.6x + 66.38$$

Założenia

Zakładamy liniowość w grupie eksperymentalnej i kontrolnej.

Założenia

Zakładamy liniowość w grupie eksperymentalnej i kontrolnej.

$$y^C = X^C \beta^C + \varepsilon^C$$
$$y^T = X^T \beta^T + \varepsilon^T = X^T \beta^C + X^T \beta^U + \varepsilon^T$$

Założenia

Zakładamy liniowość w grupie eksperymentalnej i kontrolnej.

$$y^C = X^C \beta^C + \varepsilon^C$$
$$y^T = X^T \beta^T + \varepsilon^T = X^T \beta^C + X^T \beta^U + \varepsilon^T$$

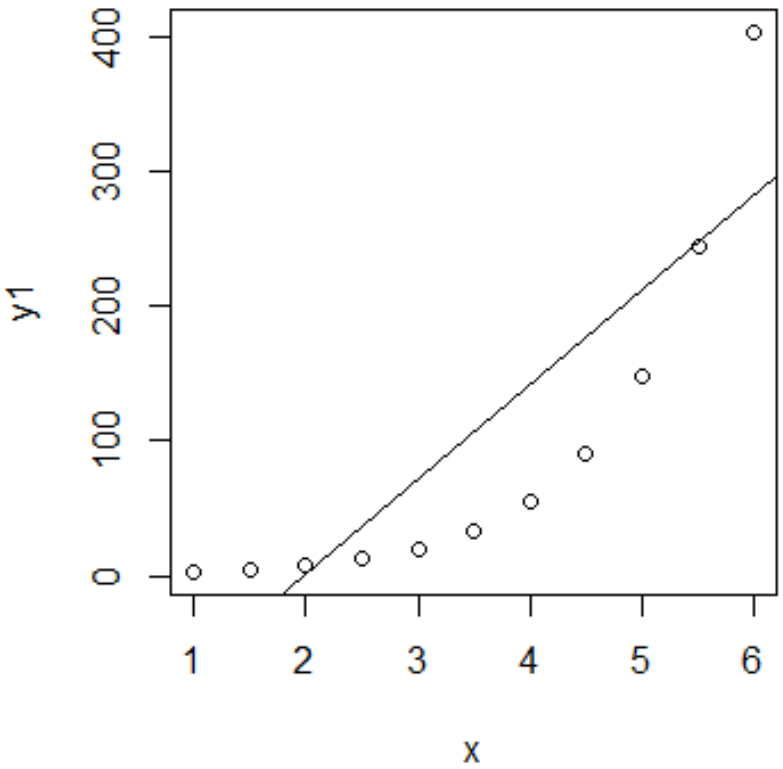
- $\mathrm{E}\, \varepsilon_i^T = 0$ i $\mathrm{Var}\, \varepsilon_i^T = \sigma^T$
- $\mathrm{E}\, \varepsilon_i^C = 0$ i $\mathrm{Var}\, \varepsilon_i^C = \sigma^C$
- Dla uproszczenia zakładamy równoliczność grup $n^T = n^C = \frac{n}{2}$

Problemy

- Założenie o liniowości jest dosyć restrykcyjne.

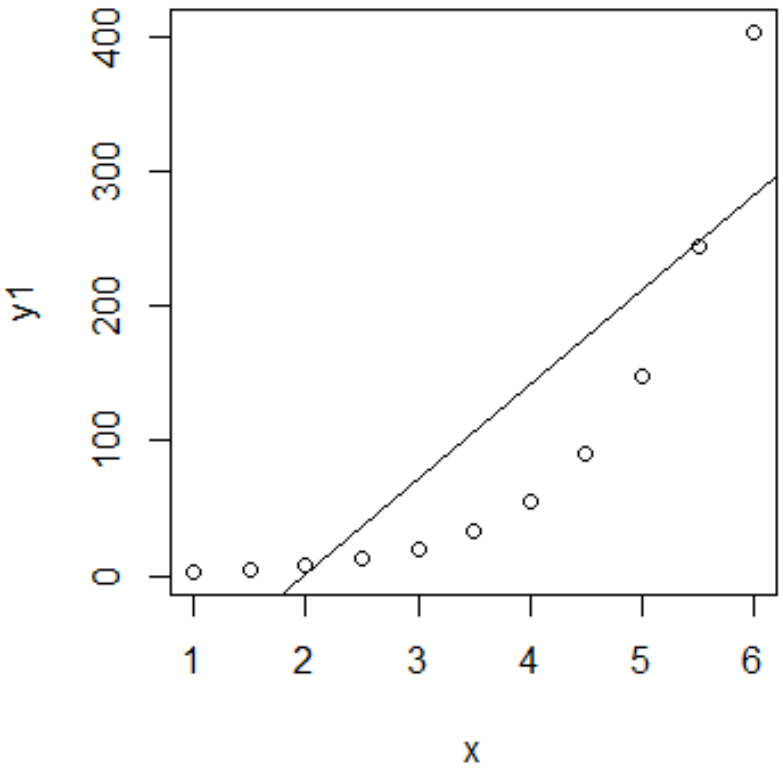
Problemy

- Założenie o liniowości jest dosyć restrykcyjne.



Problemy

- Założenie o liniowości jest dosyć restrykcyjne.



- Co jeśli cechy są bardzo mocno skorelowane?

Korelacja

- Niech $x_{.j}$ oznacza j-tą kolumnę macierzy X.

Korelacja

- Niech $x_{.j}$ oznacza j-tą kolumnę macierzy X.
- Załóżmy, że $x_{.j} = x_{.k}$, wówczas $cor(x_{.j}, x_{.k}) = 1$.

Korelacja

- Niech $x_{.j}$ oznacza j-tą kolumnę macierzy X.
- Załóżmy, że $x_{.j} = x_{.k}$, wówczas $cor(x_{.j}, x_{.k}) = 1$.
- Załóżmy, że dobre są współczynniki $\beta_j = 3$ i $\beta_k = -2$.

Korelacja

- Niech $x_{.j}$ oznacza j-tą kolumnę macierzy X.
- Załóżmy, że $x_{.j} = x_{.k}$, wówczas $cor(x_{.j}, x_{.k}) = 1$.
- Załóżmy, że dobre są współczynniki $\beta_j = 3$ i $\beta_k = -2$.
- Ale równie dobre są $\beta_j = -50000$ i $\beta_k = 50001$.

Korelacja

- Niech $x_{.j}$ oznacza j-tą kolumnę macierzy X.
- Załóżmy, że $x_{.j} = x_{.k}$, wówczas $cor(x_{.j}, x_{.k}) = 1$.
- Załóżmy, że dobre są współczynniki $\beta_j = 3$ i $\beta_k = -2$.
- Ale równie dobre są $\beta_j = -50000$ i $\beta_k = 50001$.
- Remedium: wyrzucenie zmiennych zależnych od innych zmiennych.

Cel

- Estymacja β^U.

Cel

- Estymacja β^U.
- Bierzemy nową obserwację x_{test}.

Cel

- Estymacja β^U.
- Bierzemy nową obserwację x_{test}.
- $x_{test}\hat{\beta^U}$ daje nam prognozę zysku z podjętej akcji na obserwacji x_{test}.

Estymator podwójny

Budujemy estymatory na dwóch grupach oddzielnie:

$$\hat{\beta}^T = (X^{T'} X^T)^{-1} X^T y^T$$

$$\hat{\beta}^C = (X^{C'} X^C)^{-1} X^C y^C$$

Estymator podwójny

Budujemy estymatory na dwóch grupach oddzielnie:

$$\hat{\beta}^T = (X^{T\prime} X^T)^{-1} X^T y^T$$

$$\hat{\beta}^C = (X^{C\prime} X^C)^{-1} X^C y^C$$

Estymator podwójny

$$\hat{\beta}_d^U = \hat{\beta}^T - \hat{\beta}^C$$

Estymator różnicowy

Czy da się estymować β^U w jednym kroku?

Estymator różnicowy

Czy da się estymować β^U w jednym kroku?

Estymator różnicowy

Oznaczmy:

$$\tilde{y}_i = \begin{cases} 2y_i^T & \text{if } g_i = T, \\ -2y_i^C & \text{if } g_i = C. \end{cases}$$

Estymator różnicowy ma następującą postać:

$$\hat{\beta}_r^U = (X'X)^{-1}X'\tilde{y}$$

Jak mierzyć skuteczność modelu?

- Chcemy porównać prawdziwą odpowiedź z naszą prognozą.

Jak mierzyć skuteczność modelu?

- Chcemy porównać prawdziwą odpowiedź z naszą prognozą.
- $MSE = \mathrm{E}\,||y^U_{test} - X_{test}\hat{\beta^U}||^2$.

Jak mierzyć skuteczność modelu?

- Chcemy porównać prawdziwą odpowiedź z naszą prognozą.
- $MSE = \mathrm{E}\,||y^U_{test} - X_{test}\hat{\beta^U}||^2$.
- MSE jest zależne od obciążenia i wariancji estymatora.

Jak mierzyć skuteczność modelu?

- Chcemy porównać prawdziwą odpowiedź z naszą prognozą.
- $MSE = \mathrm{E}\,||y^U_{test} - X_{test}\hat{\beta^U}||^2$.
- MSE jest zależne od obciążenia i wariancji estymatora.
- Czy wprowadzając wzór na MSE nie oszukuję?.

Własności estymatorów

- $\hat{\beta}_d^U$ i $\hat{\beta}_r^U$ są nieobciążone.

Własności estymatorów

- $\hat{\beta}_d^U$ i $\hat{\beta}_r^U$ są nieobciążone.
- Oznaczmy $\beta^* = \beta^T + \beta^C$:
- Gdy $\beta^* >> 0$ i $n >> p$:

$$\operatorname{Var} \hat{\beta}_d^U < \operatorname{Var} \hat{\beta}_r^U$$

Własności estymatorów

- $\hat{\beta}_d^U$ i $\hat{\beta}_r^U$ są nieobciążone.
- Oznaczmy $\beta^* = \beta^T + \beta^C$:
- Gdy $\beta^* >> 0$ i $n >> p$:

$$\operatorname{Var} \hat{\beta}_d^U < \operatorname{Var} \hat{\beta}_r^U$$

- Gdy $\beta^* \approx 0$ lub p nie jest mocno mniejsze od n:

$$\operatorname{Var} \hat{\beta}_d^U > \operatorname{Var} \hat{\beta}_r^U$$

Własności estymatorów

- $\hat{\beta}_d^U$ i $\hat{\beta}_r^U$ są nieobciążone.
- Oznaczmy $\beta^* = \beta^T + \beta^C$:
- Gdy $\beta^* >> 0$ i $n >> p$:

$$\operatorname{Var} \hat{\beta}_d^U < \operatorname{Var} \hat{\beta}_r^U$$

- Gdy $\beta^* \approx 0$ lub p nie jest mocno mniejsze od n:

$$\operatorname{Var} \hat{\beta}_d^U > \operatorname{Var} \hat{\beta}_r^U$$

- Czy jestem w stanie stworzyć model, który połączy zalety obydwu?

Estymator korygowany

1. Wyestymujmy β^*:

$$\hat{\beta}^* = (X'X)^{-1}X'y$$

Estymator korygowany

1. Wyestymujmy β^*:

$$\hat{\beta}^* = (X'X)^{-1}X'y$$

2. Skorygujmy y

$$y^{corr} = y - X\hat{\beta}^*$$

Estymator korygowany

1. Wyestymujmy β^*:
$$\hat{\beta}^* = (X'X)^{-1}X'y$$
2. Skorygujmy y
$$y^{corr} = y - X\hat{\beta}^*$$
3. Budujemy estymator różnicowy na y^{corr}:
$$\hat{\beta}^U_{corr} = (X'X)^{-1}X'\widetilde{y^{corr}}$$

Estymator korygowany

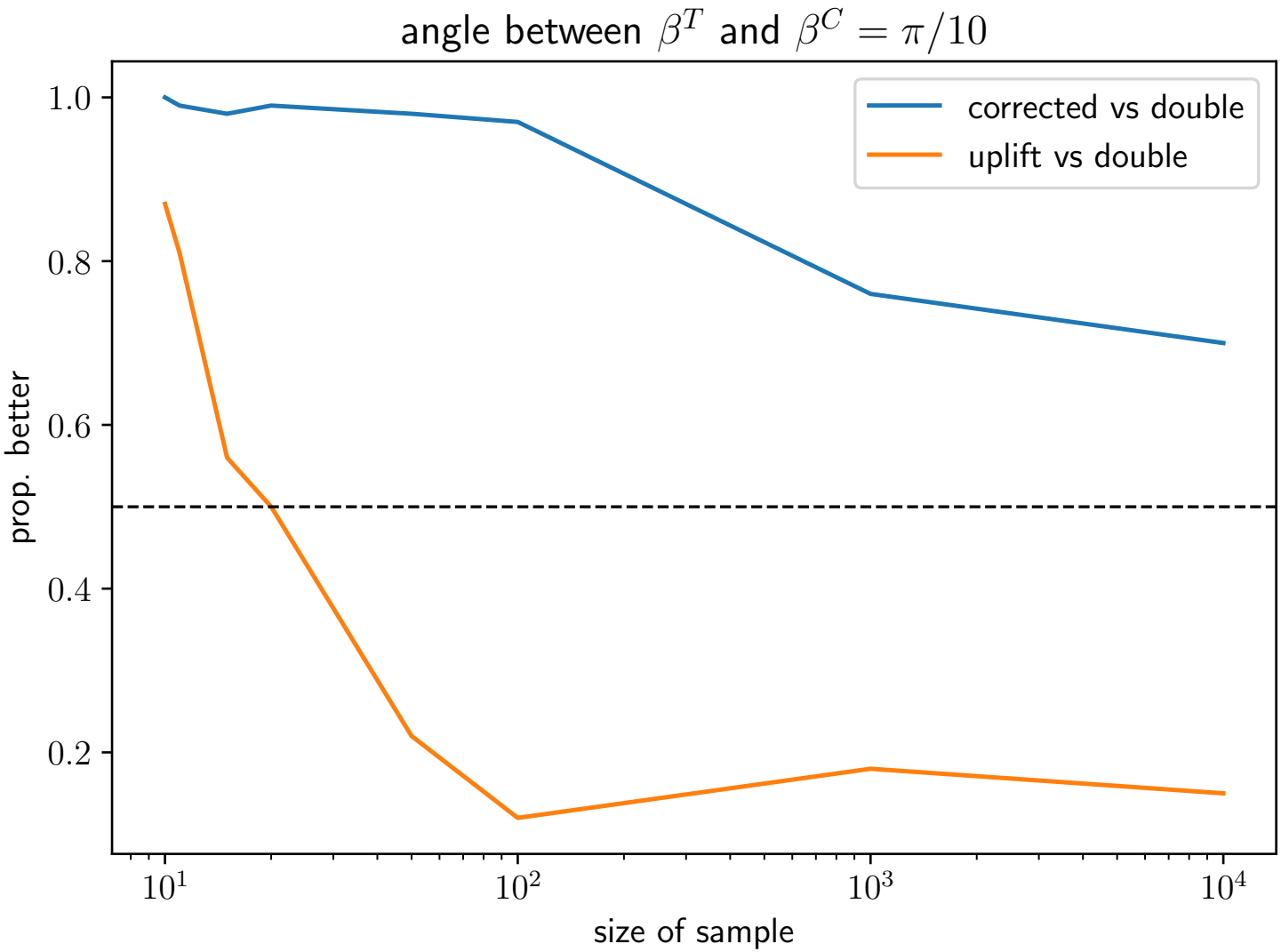
1. Wyestymujmy β^*:
$$\hat{\beta}^* = (X'X)^{-1}X'y$$
2. Skorygujmy y
$$y^{corr} = y - X\hat{\beta}^*$$
3. Budujemy estymator różnicowy na y^{corr}:
$$\hat{\beta}^U_{corr} = (X'X)^{-1}X'\widetilde{y^{corr}}$$
4. Pozbywamy się wrażliwości na β^*.

Estymator korygowany

1. Wyestymujmy β^*:
$$\hat{\beta}^* = (X'X)^{-1}X'y$$
2. Skorygujmy y
$$y^{corr} = y - X\hat{\beta}^*$$
3. Budujemy estymator różnicowy na y^{corr}:
$$\hat{\beta}^U_{corr} = (X'X)^{-1}X'\widetilde{y^{corr}}$$ K.Rudaś, S.Jaroszewicz 2018
4. Pozbywamy się wrażliwości na β^*.

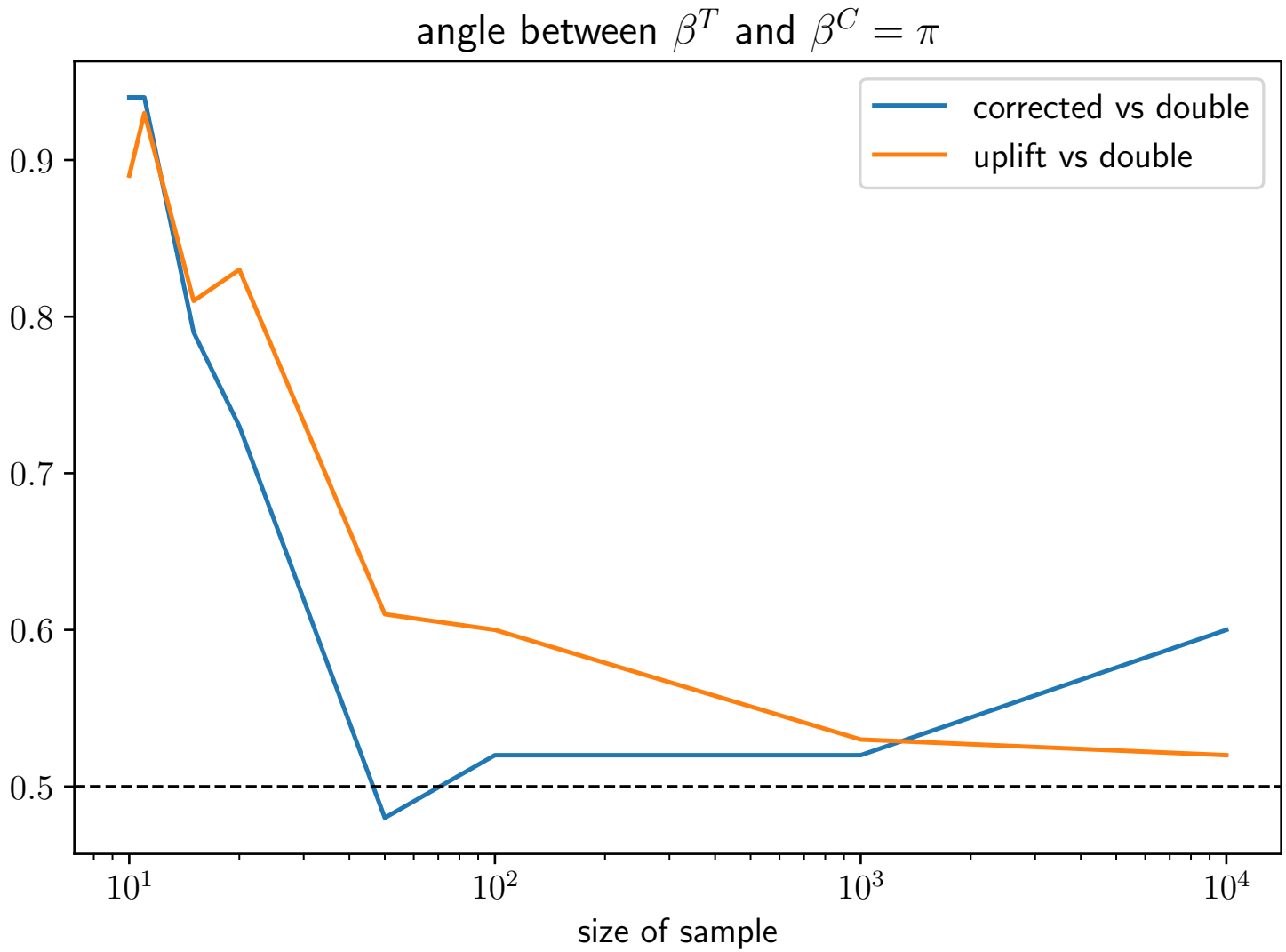
Porównanie modeli

Kąt między β^T i β^C równy $\frac{\pi}{10}$.



Porównanie modeli

Kąt między β^T i β^C równy π.



Błąd

- $\hat{\beta}^U_{corr}$ ma bardzo niskie MSE,

Błąd

- $\hat{\beta}^U_{corr}$ ma bardzo niskie MSE,
- To znaczy, że $\mathrm{Var}\,\hat{\beta}^U_{corr}$ i obciążenie małe,

Błąd

- $\hat{\beta}^U_{corr}$ ma bardzo niskie MSE,
- To znaczy, że $\mathrm{Var}\,\hat{\beta}^U_{corr}$ i obciążenie małe,
- $\hat{\beta}^U_{corr}$ to drobna modyfikacja $\hat{\beta}^U_r$, więc pewnie nieobciążony.

Błąd

- $\hat{\beta}^U_{corr}$ ma bardzo niskie MSE,
- To znaczy, że $\mathrm{Var}\,\hat{\beta}^U_{corr}$ i obciążenie małe,
- $\hat{\beta}^U_{corr}$ to drobna modyfikacja $\hat{\beta}^U_r$, więc pewnie nieobciążony.
- W rzeczywistości tak nie jest!

Inne podejścia

- Znalezienie "przełącznika"przestawiającego $\hat{\beta}_r^U$ na $\hat{\beta}_d^U$.

Inne podejścia

- Znalezienie "przełącznika"przestawiającego $\hat{\beta}_r^U$ na $\hat{\beta}_d^U$.
- Ściąganie estymatora $\hat{\beta}_d^U$.

Inne podejścia

- Znalezienie "przełącznika"przestawiającego $\hat{\beta}_r^U$ na $\hat{\beta}_d^U$.
- Ściąganie estymatora $\hat{\beta}_d^U$.
- Regularyzacja $\hat{\beta}_d^U$.

Procedura testowa

- W rzeczywistości mamy tylko połowę informacji dla każdej obserwacji:

Procedura testowa

- W rzeczywistości mamy tylko połowę informacji dla każdej obserwacji:
- Nie możemy porównać: $X_i\hat{\beta^U}$ z $y_i^T - y_i^C$:

Procedura testowa

- W rzeczywistości mamy tylko połowę informacji dla każdej obserwacji:
- Nie możemy porównać: $X_i\hat{\beta^U}$ z $y_i^T - y_i^C$:
- Aby móc uzyskać wyniki podobne do MSE używamy miary ATT:

$$\frac{1}{n^T}\sum_{i=1}^{n^T} X_i\hat{\beta^U} \text{ z } \frac{1}{n^T}\sum_{i=1}^{n^T} y_i^T - \frac{1}{n^C}\sum_{i=1}^{n^C} y_i^C$$

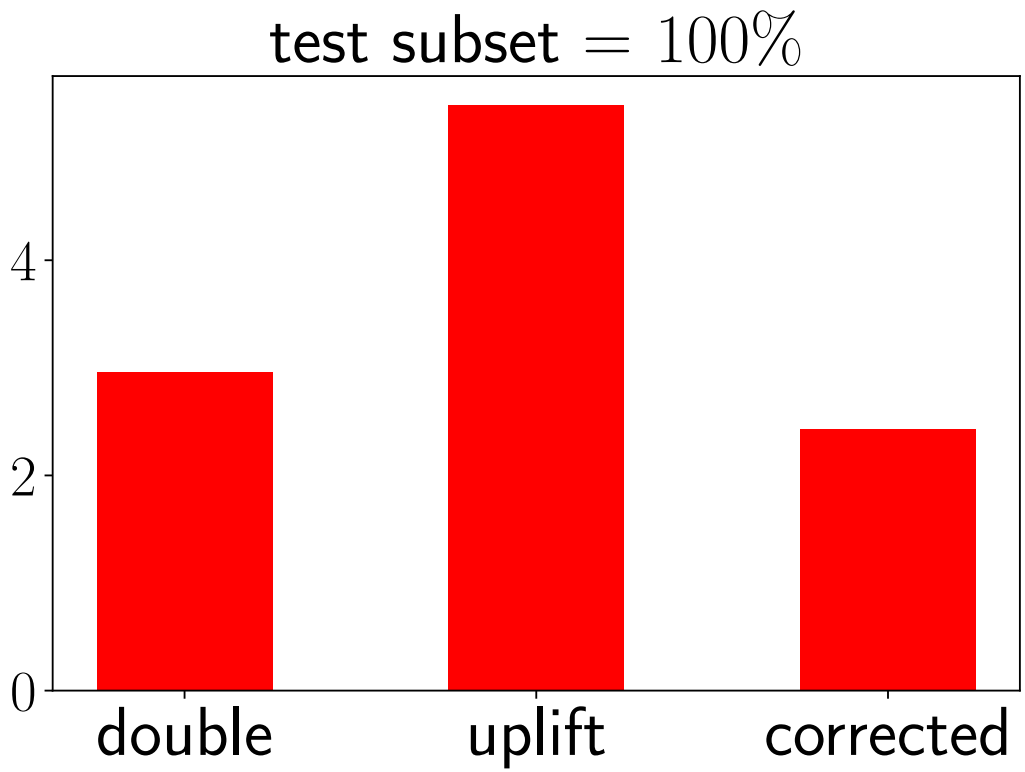
Dane rzeczywiste - OSNAP

- OSNAP - kampania, mająca na celu popularyzację aktywnego stylu życia wśród dzieci w USA.

Dane rzeczywiste - OSNAP

- OSNAP - kampania, mająca na celu popularyzację aktywnego stylu życia wśród dzieci w USA.
- Cel: znaleźć dzieci, dla których skierowanie kampanii przyniesie najlepsze rezultaty.

Rezultaty



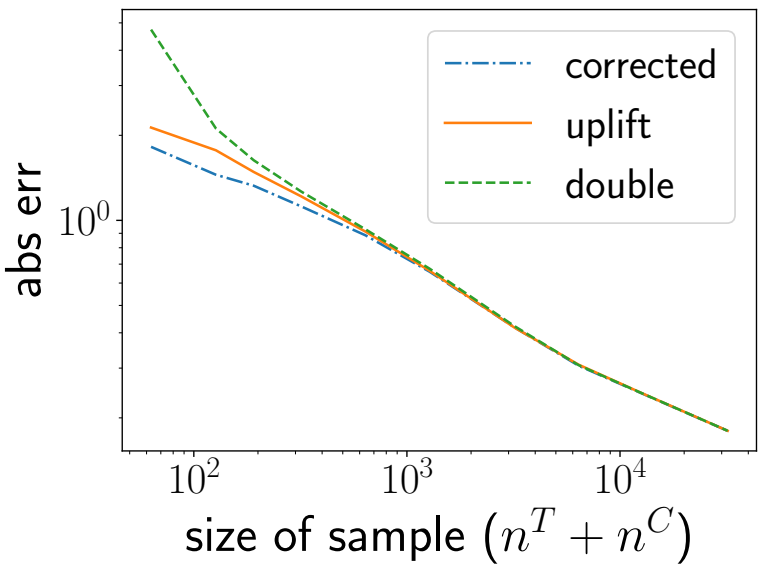
Dane rzeczywiste - Hillstrom

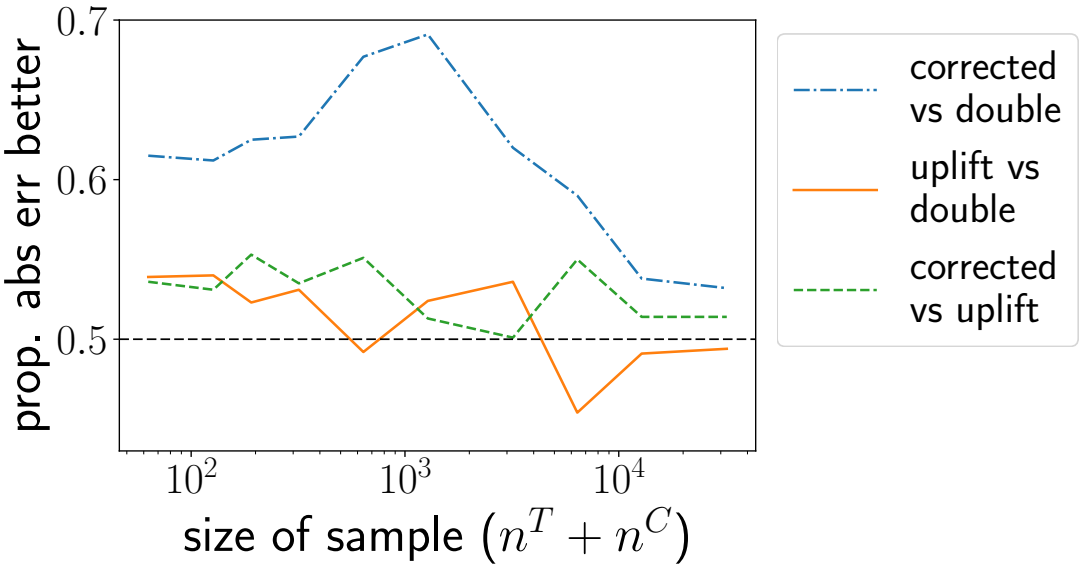
- Hillstrom - kampania e-mailowa, mająca zachęcać klientów do kupowania przez internet,

Dane rzeczywiste - Hillstrom

- Hillstrom - kampania e-mailowa, mająca zachęcać klientów do kupowania przez internet,
- Cel: znaleźć tych klientów, którzy kupowali więcej, ponieważ kampania do nich dotarła.

Rezultaty


abs err
10^0
10^2
10^3
10^4
size of sample $(n^T + n^C)$
corrected
uplift
double


prop. abs err better
0.7
0.6
0.5
10^2
10^3
10^4
size of sample $(n^T + n^C)$
corrected vs double
uplift vs double
corrected vs uplift

- Modelowanie przyczynowości jest możliwe, ale pod pewnymi warunkami,

- Modelowanie przyczynowości jest możliwe, ale pod pewnymi warunkami,
- Istnieją modele, które pozwalają określenie które czynniki stają się istotne kiedy akcja została podjęta.

- Modelowanie przyczynowości jest możliwe, ale pod pewnymi warunkami,
- Istnieją modele, które pozwalają określenie które czynniki stają się istotne kiedy akcja została podjęta.
- Witaminy nie leczą raka!!!.

Dziękuję za uwagę