Zaawansowana analityka biznesowa – siła modeli predykcyjnych

226160-D

Imputacja danych brakujących w badaniu prospektywnym w charakterze obserwacji cechy w dwóch momentach

Weronika Banasiak

Numer albumu 72698

Joanna Hawrysz

Numer albumu 122952

Ksenia Soroka

Numer albumu 89716

Joanna Zakęs

Numer albumu 100930

WARSZAWA 2023

Spis treści

1. Cel projektu

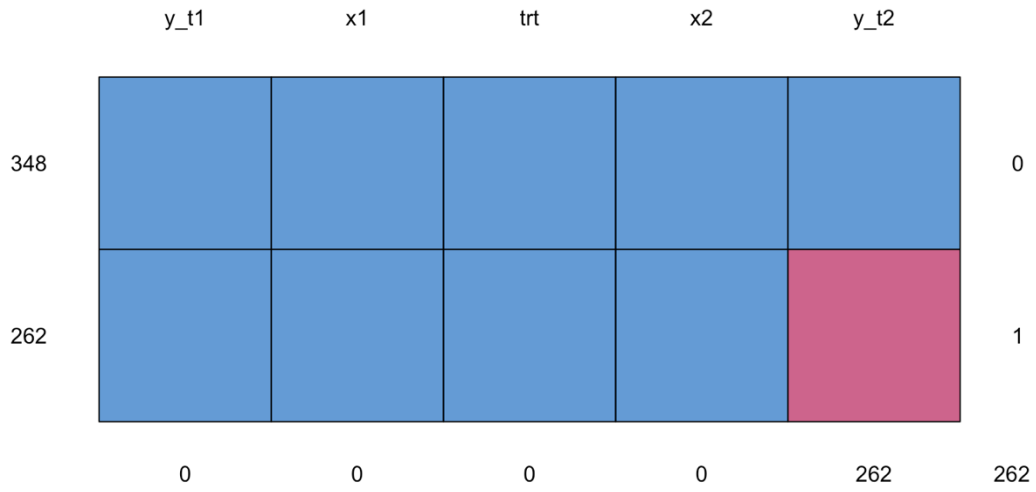
Celem projektu było wyznaczenie średniej wielkości zmiany w podgrupach ze względu na wartość zmiennej trt i ocena różnicy pomiędzy tymi grupami.

Do celów szczegółowych zadania zaliczono:

1. Ustalenie wielkości średniej zmiany w zależności dla grup wystawionych i niewystawionych na działanie czynnika;
2. Poznanie rozkładu zmiennych i ich analizę graficzną;
3. Ocenę zależności pomiędzy prawdopodobieństwem braku danych a pozostałymi cechami;
4. Oszacowanie modelu analizy kowariancji MCAR oraz MAR wraz z oceną zasadności stosowania założenia o nielosowym charakterze braków danych MNAR i wykonanie *tipping point analysis*;
5. Podsumowanie i ocenę wyników estymacji modelu właściwego w kontekście braków danych.

2. Eksploracyjna analiza danych

W pierwszej kolejności, w celu zapoznania się ze strukturą zbioru danych sprawdzono wzorzec braków danych (Rys. 1) oraz podstawowe miary statystyczne dla zmiennych ciągłych (Rys.2). Zaobserwowano, że braki występują jedynie dla zmiennej y_t2 w liczbie 262 co odpowiada w zaokrągleniu 43% wszystkich obserwacji.



Rysunek 1 Rozkład braków danych w zbiorze

W statystykach opisowych w szczególności zwrócono uwagę na to, jakie różnice występują pomiędzy dwoma pomiarami zmiennej y (y_t1 i y_t2). O ile średnia i mediana dla obu z nich mają zbliżone wartości, o tyle y_t2 ma mniejsze minimum i większe maksimum (co oznacza również większy rozstęp).

```
      y_t1             x1            y_t2
 Min.   : 6.947   Min.   :34.55   Min.   : 4.388
 1st Qu.:10.308   1st Qu.:45.40   1st Qu.: 9.452
 Median :10.994   Median :49.75   Median :10.902
 Mean   :10.959   Mean   :49.64   Mean   :10.912
 3rd Qu.:11.622   3rd Qu.:53.19   3rd Qu.:12.453
 Max.   :14.538   Max.   :67.68   Max.   :17.207
                                  NA's   :262
```

Rysunek 2 Podstawowe miary statystyczne dla zmiennych ciągłych

Dla tych samych zmiennych sprawdzono również wariancję, skośność i kurtozę (Tabela 1). Nie zauważono znaczących różnic poza wzrostem wariancji pomiędzy y_t1 a y_t2, co jest naturalną implikacją wartości maksimum i minimum opisanych w poprzednim akapicie.

	y_t1	x1	y_t2
Wariancja	0.9955524	5.8710370	2.398095
Skośność	-0.05410897	0.08763251	-0.08162821
Kurtoza	3.265327	3.140376	2.961661

Tabela 1 Wariancja, skośność, kurtoza dla zmiennych ciągłych

Dla zmiennych skokowych sprawdzono liczności dla każdej wartości (Tabela 2).

Zmienna	x2		trt	
Wartość	0	1	1	2
Liczność	306	304	300	310

Tabela 2 Liczności dla zmiennych skokowych

Proporcje kategorii zmiennych oscylują wokół 50%, co oznacza, że nie występuje problem zbyt małej kategorii zmiennych. Dla wszystkich zmiennych wykonano również macierz korelacji (Rys. 3), przede wszystkim, żeby sprawdzić czy którąś z nich można potencjalnie użyć do imputacji braków danych w y_t2, jeśli taka imputacja będzie wskazana. Nie zauważono znaczącego wpływu. Warto zaznaczyć, że jeśli istniałaby potrzeba mierzenia siły wpływu pomiędzy zmiennymi skokowymi takimi jak x2 i trt, należałoby użyć asocjacji, jednakże nie jest to celem tego opracowania. Warto zauważyć, że zmienna y_t2 największą korelację posiada ze zmienną y_t1 (0.27), co oznacza, że oczekujemy słabej relacji dodatniej.

	y_t1	x1	trt	x2	y_t2
y_t1	1.000000000	-0.01296480	-0.03597101	-0.006452217	0.26602187
x1	-0.012964802	1.00000000	0.09925464	-0.019342099	0.03046793
trt	-0.035971012	0.09925464	1.00000000	-0.517275551	-0.25741802
x2	-0.006452217	-0.01934210	-0.51727555	1.000000000	0.05918109
y_t2	0.266021866	0.03046793	-0.25741802	0.059181095	1.00000000

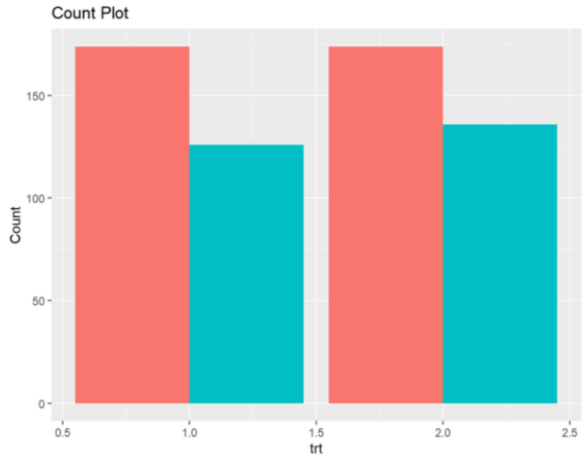
Rysunek 3 Analiza korelacji dla zmiennych

Następnie analogicznie do poprzedniego akapitu, wykonano analizę kowariancji (Rys. 4). Ponownie nie zauważono na tyle dużych wartości, aby mogły sugerować jakimi zmiennymi można potencjalnie imputować braki w y_t2 (chociaż warto zwrócić uwagę na wartość 0,44 dla x1 i y_t2).

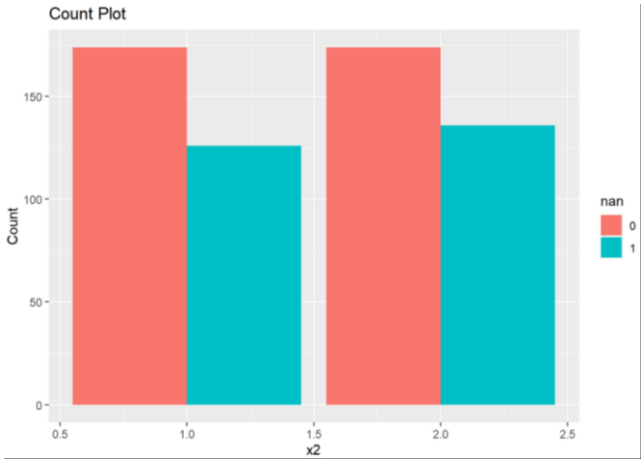
	y_t1	x1	trt	x2	y_t2
y_t1	0.980302113	-0.07657654	-0.01783313	-0.003198565	0.63163143
x1	-0.076576545	35.58769336	0.29648016	-0.057772308	0.43587230
trt	-0.017833128	0.29648016	0.25072046	-0.129682997	-0.30910090
x2	-0.003198565	-0.05777231	-0.12968300	0.250687336	0.07105843
y_t2	0.631631429	0.43587230	-0.30910090	0.071058432	5.75086076

Rysunek 4 Analiza kowariancji dla zmiennych

Ponieważ nie znaleziono interesujących zależności w statystykach opisowych, przystąpiono do analizy graficznej danych. W celu czytelności i poprawy możliwości wnioskowania na podstawie wykresów dodano zmienną m, która przyjmuje wartość 0, jeśli w kolumnie y_t2 nie było braku danych i 1 jeśli był. Na podstawie wykresów 1 i 2 sprawdzono, czy w zmiennych skokowych liczby braków danych w zależności od wartości zmiennej różnią się – nie zanotowano niepokojących różnic.

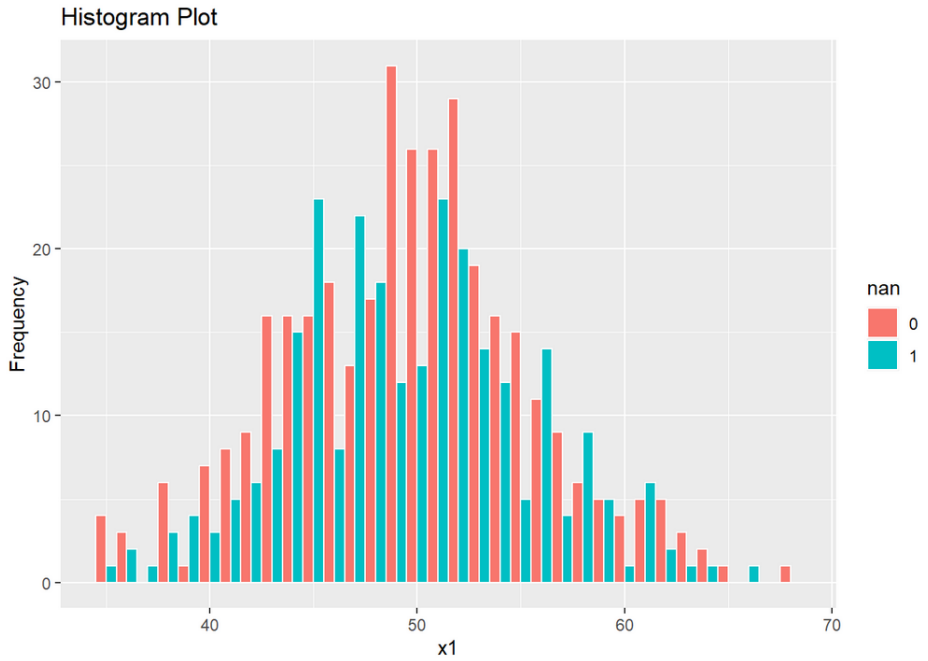


Wykres 1 Liczności dla poszczególnych wartości zmiennej trt w zależności od braków danych

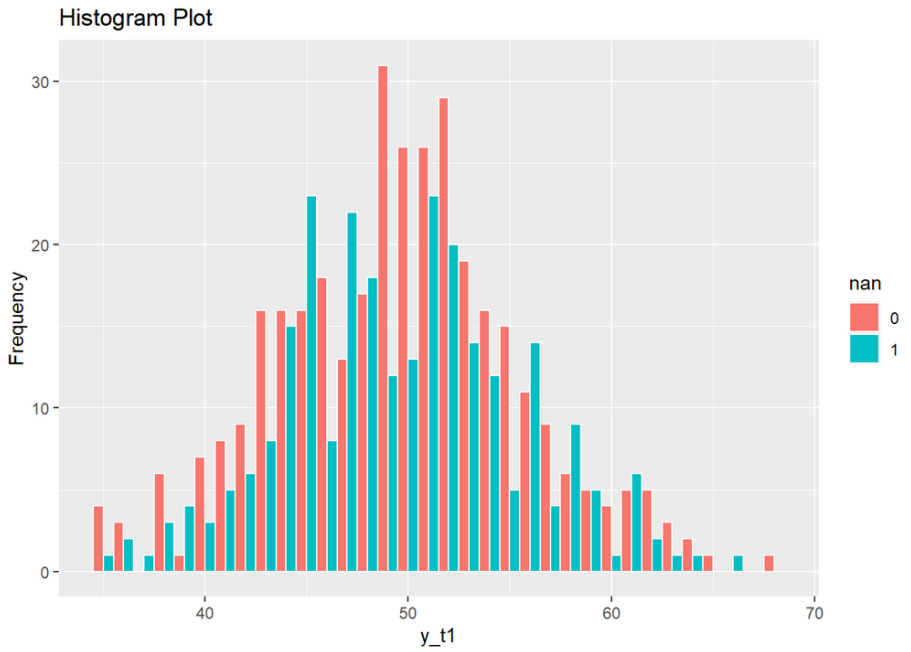


Wykres 2 Liczności dla poszczególnych wartości zmiennej x2 w zależności od braków danych

Dla zmiennych ciągłych wykonano histogramy (Wykres 3 i 4), również z opcją wyróżnienia kolorem wykresu w zależności od braków danych w y_t2. Zauważono widoczną przewagę danych kompletnych zlokalizowanych w obszarze odpowiadającym wartościom x1 zbliżonym do średniej – w najliczniejszych binsach, natomiast w przypadku braku danych, widoczny jest wzrost w okolicach wartości 45 i 48 by następnie zaobserwować spadek i kolejny znaczący wzrost dla wartości 51 i 52.

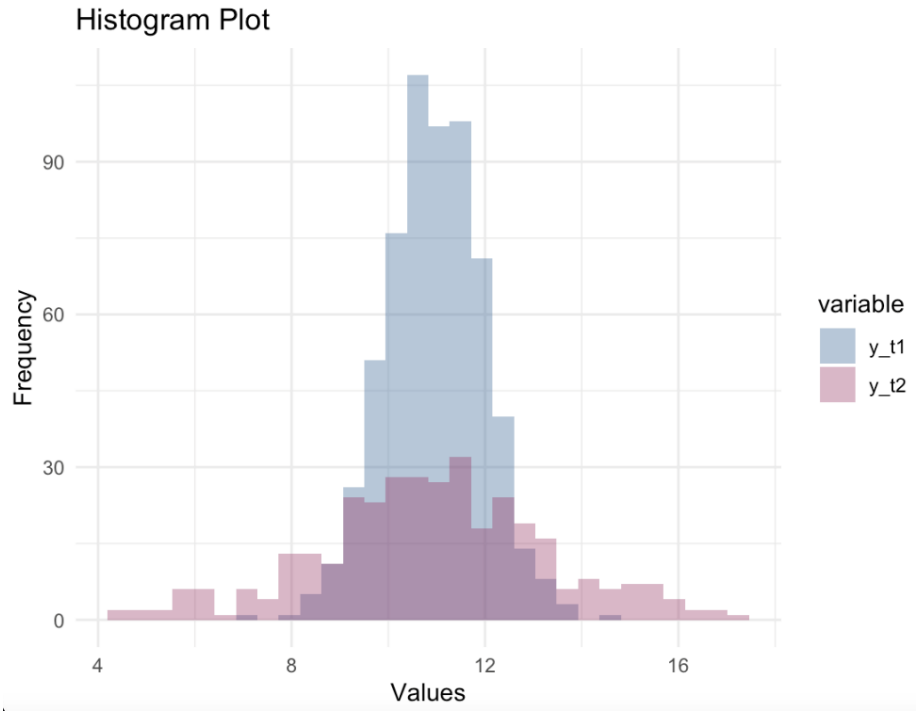


Wykres 3 Histogram zmiennej x1 w zależności od braków danych



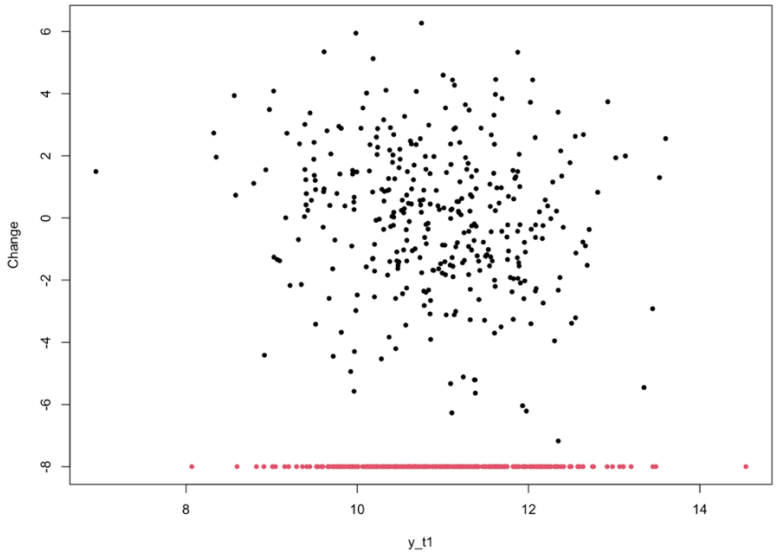
Wykres 4 Histogram zmiennej y_t1 w zależności od braków danych

Porównano również histogramy dla y_t1 i y_t2 (Wykres 5). Oba wykresy mają podobną symetrię, natomiast dla zmiennej y_t2 odnotowano wartości z zakresu od 4 do 8 oraz od 14 do 16, które prawie nie występują dla zmiennej y_t1. Pomimo, że wartości między 10 a 12 są dla zmiennej y_t2 najliczniejszymi, są one znacznie mniej liczne niż dla zmiennej y_t1.

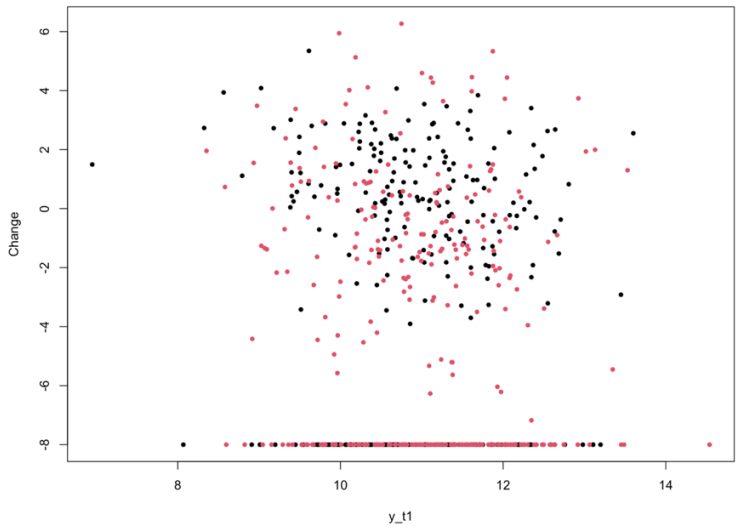


Wykres 5 Wspólny histogram zmiennych y_t1 i y_t2

Zwizualizowano również związek między zmianą (change) a zmienną y_t1. Przedstawiono braki danych poprzez umieszczenie ich na poziomej linii o wartości -8. Zabieg ten miał na celu umożliwić lepsze odczytanie wykresu. Na wykresie 6 kolor punktów zostaje zdefiniowany przez wartości kategorialne z kolumny m, zaś na wykresie 7 z kolumny trt. Na wykresach nie widać, aby zmienna trt poddana była jakiemuś schematowi w połączeniu ze zmiennymi y_t1 lub change.

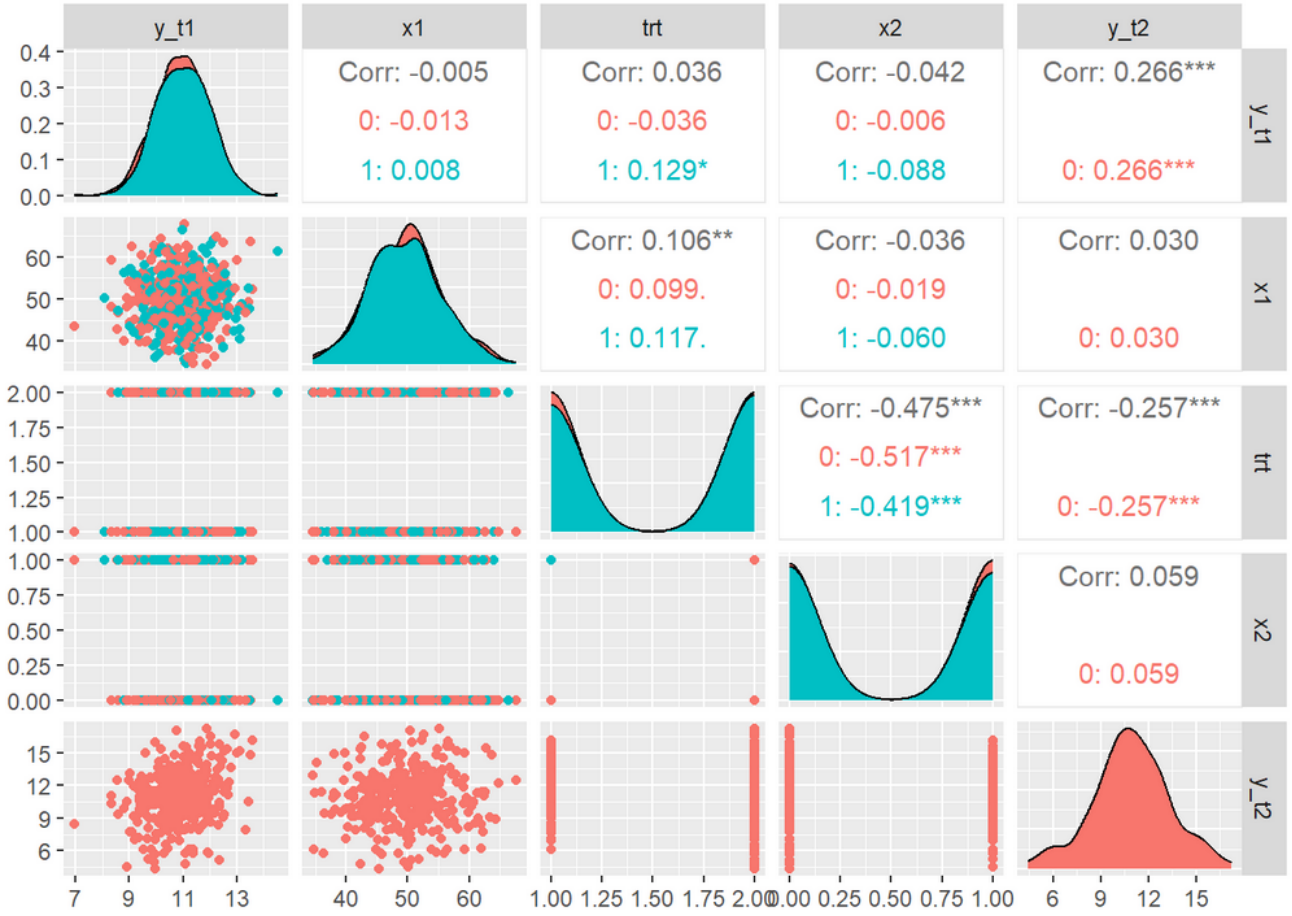


Wykres 6 Zależność między zmianą a y_t1 w podziale na grupy z brakami danych i bez nich



Wykres 7 Zależność między zmianą a y_t1 w podziale na kategorie zmiennej trt

W celu sprawdzenia czy na wykresach par zmiennych będą widoczne jakieś zależności pomiędzy wartościami a brakami danych wykonano również siatkę, która zawiera wykresy par dla wszystkich możliwych kombinacji zmiennych (Wykres 8). Nie zauważono znaczących różnic, które mogłyby sugerować, że braki danych są nielosowe.



Wykres 8 Wykres par zmiennych w zależności od braków danych

Zmienne poddano analizie wartości odstających oraz analizie różnic pomiędzy grupą z brakami danych dla zmiennej y_t2 a grupą bez braków danych. Do oceny obecności wartości odstających wykorzystano testy Grubbsa i Rosnera. Wyniki przedstawiono w Tabelach 3 i 4.

Zmienna	**G**	**U**	**p-value**
Y_t1 lewy ogon rozkładu	4.02978	0.97329	0.01526
Y_t1 prawy ogon rozkładu	3.59474	0.97875	0.09247
Y_t2 lewy ogon rozkładu	2.72077	0.97861	1
Y_t2 prawy ogon rozkładu	2.62487	0.98009	1
x1 lewy ogon rozkładu	3.07393	0.98446	0.6215

x2 prawy ogon rozkładu	2.56906	0.98914	1

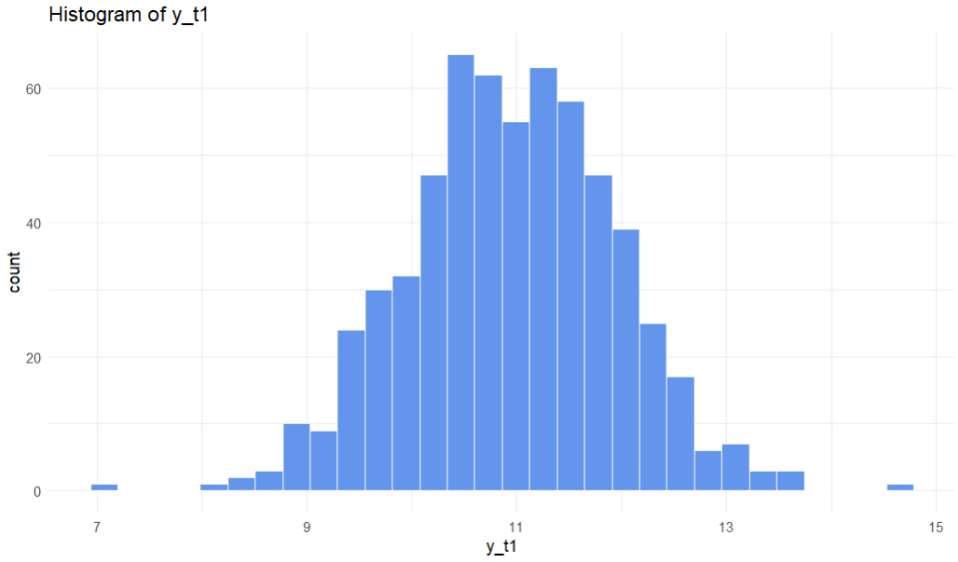
Tabela 3 Wyniki testu Grubbsa

Zmienna	Liczba wykrytych wartości odstających
y_t1	1
y_t2	0
x1	0

Tabela 4 Wyniki testu Rosnera

Oba testy wskazują istnienie jednej wartości odstającej dla zmiennej y_t1, będącej minimum tej zmiennej w zbiorze, przy poziomie istotności 0,05. Nie wskazują natomiast zmiennych odstających w prawym ogonie rozkładu tej zmiennej ani w rozkładach zmiennych y_t2 oraz x1.

Na podstawie Wykresu 9 zaobserwowano, że minimum zmiennej y_t1 w zbiorze nie wydaje się być skrajną i nieprawdopodobną obserwacją. Istnieją podejrzenia, że jest to realistyczna wartość pochodząca z rzeczywistego procesu generującego dane. W związku z tym obserwacja ta nie została usunięta ze zbioru.



Wykres 9 Histogram zmiennej y_t1

W Tabelach 5 i 6 przedstawiono wyniki testu t dla zmiennych y_t1 i x1 oraz testu chi-kwadrat Pearsona dla zmiennych x2 i trt. Porównano grupę z brakami danych dla zmiennej y_2 do grupy bez braków danych. Zarówno dla zmiennej y_t1, jak i zmiennej x1 brak podstaw do odrzucenia hipotezy zerowej

o braku różnicy średnich między grupami, przy poziomie istotności 0,05. Podobnie dla zmiennych dyskretnych, w obu przypadkach brak podstaw do odrzucenia hipotezy zerowej o braku związku zmiennych z brakami danych, przy poziomie istotności 0,05. Braki danych dla zmiennej y_t2 nie są związane z żadną z pozostałych zmiennych.

Zmienna	t	p-value
y_t1	-1.3193	0.1876
x1	0.30122	0.7634

Tabela 5 Wyniki testu t

Zmienna	$\chi 2$	p-value
x2	0.17683	0.6741
trt	0.2178	0.6407

Tabela 6 Wyniki testu chi-kwadrat

3. Imputacja danych

3.1. Imputacja przy założeniach MCAR

W dalszej części pracy oszacowano modele analizy kowariancji przy założeniach MCAR oraz MAR. W pierwszej kolejności skupiono się na brakach danych typu „*Missing Completely at Random*". Ustawiono referencyjną kategorię dla zmiennej trt jako „2" (grupa kontrolna) oraz przekonwertowano ją na zmienną kategoryczną. Ustawiono również kontrasty stosowane w modelach statystycznych:

- contr.treatment – podejście, które traktuje pierwszy poziom jako referencyjny, a następnie porównuje każdy kolejny poziom z referencyjnym. W tym przypadku, po użyciu *relevel*, referencyjnym poziomem będzie ten, który wcześniej był drugim poziomem.
- contr.poly – podejście to stosuje kontrast wielomianowy, który może być używany w analizie regresji wielomianowej.

Ponieważ braki danych typu MCAR nie wpływają na rozkład zmiennej istnieje możliwość zastosowania metody CC (*complete-case analysis*), która zakłada wykorzystanie tylko kompletnych obserwacji. Wynik modelu ANCOVA dla danych kompletnych (tj. 345 obserwacji) przedstawiono na

Rys. 5. Wynik testu F jest statystycznie istotny przy ustalonym poziomie istotności 5%, a więc przynajmniej jedna średnia różni się istotnie statystycznie od pozostałych.

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
y_t1         1   43.03  43.033  8.5738 0.003637 **
trt          1  122.74 122.744 24.4554 1.19e-06 ***
Residuals  345 1731.58   5.019
```

Rysunek 5 Wynik analizy kowariancji ANCOVA

W dalszej analizie przeprowadzono test post-hoc Tukey'a HSD (*Honestly-Significant Difference*), a jego wyniki zaprezentowano na Rys. 6.

```
Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
1 - 2 == 0   1.1886     0.2403   4.945 1.19e-06 ***
```

Rysunek 6 Wynik testu post-hoc Tukey'a

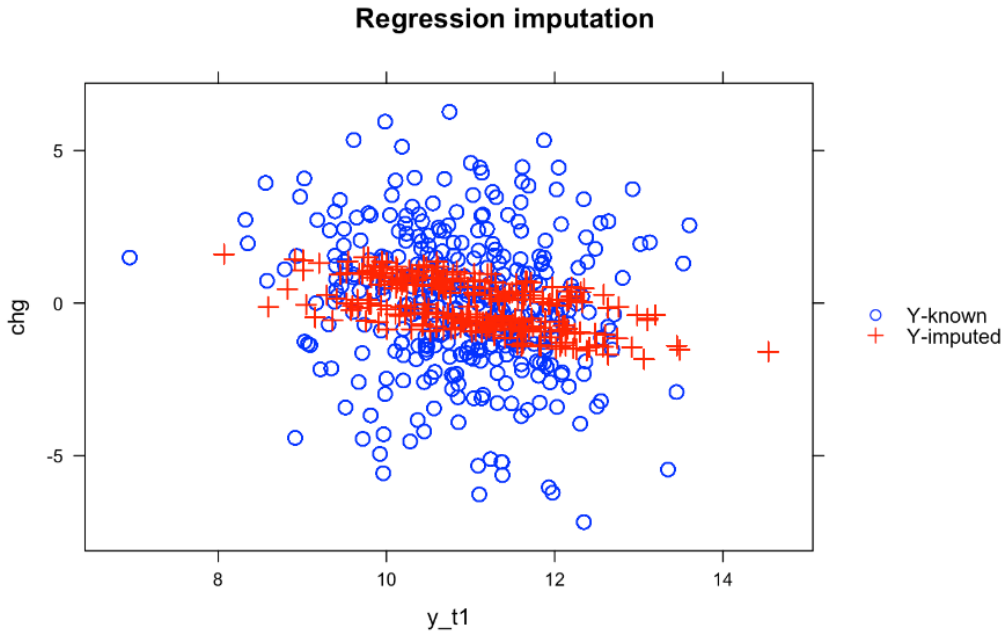
Odrzucamy hipotezę zerową mówiącą o równości średnich na rzecz hipotezy alternatywnej. Porównania post-hoc za pomocą testu Tukey'a HSD wykazały istotne statystycznie różnice między grupami ($p<0{,}001$). Grupa poddana czynnikowi trt ma przeciętnie o ok. 1,19 wyższą zmianę między momentem y_t2 a y_t1 w stosunku do grupy kontrolnej.

3.2. Imputacja przy założeniach MAR

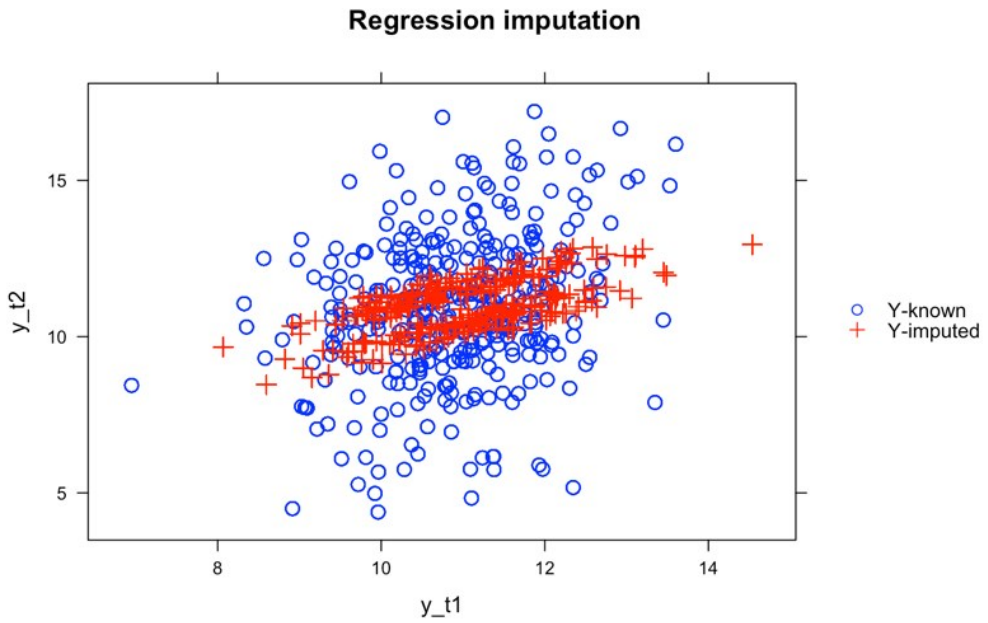
Przechodząc do braków danych typu MAR (*Missing at Random*), przeprowadzono analizę przy użyciu różnych wariantów imputacji: regresji liniowej, bayesowskiej regresji liniowej, drzewa decyzyjnego oraz imputacji „hot deck" (metoda najbliższego sąsiedztwa) dla porównania rezultatów. Ustawiono wspólne ziarno losowości równe '1234' oraz stworzono nowy zbiór: *z17new,* który nie zawierał zbędnych informacji. W związku z tym w zbiorze pozostały zmienne: y_t1, trt, x1, x2, y_t2. Ponadto, ponieważ wartości zmiennej y_t1 charakteryzowały się unikalnością, można było stworzyć zbiór *missing_vector* mapujący czy dla danej zmiennej y_t1, obserwacja y_t2 była brakująca (1) czy też nie (0). Liczba imputacji wielokrotnych w każdym z wariantów wyniosła 25. Po przeprowadzonych imputacjach wyniki zostają zapisane w nowych zbiorach nazwanych *matmi*. Tworzy się kompletny zestaw danych, w którym brakujące wartości zostały uzupełnione. Opcja „*long*" oznacza, że dane dla każdej obserwacji są reprezentowane w osobnych wierszach, nie zaś w kolumnach. Następnie zbiory *matmi* zostają nadpisane poprzez połączenie zbiorów *imp* ze zbiorem *missing_vector* na podstawie

kolumny y_t1. W ostatnim kroku obliczono zmianę (zmienna „chg") oraz przedstawiono końcowy wynik każdej imputacji na wykresie punktowym. Dla każdej metody wykonano dwa wykresy: wykres zależności zmiennej chg od y_t1 oraz zmiennej y_t2 od y_t1, z grupowaniem według zmiennej m, z podziałem na y_t2 znane (Y_known) oraz zaimputowane (Y_imputed).

Na Wykresach 10 i 11, przedstawiających imputację przy użyciu regresji liniowej, można zaobserwować, że imputacje pasują jedynie do niewielkiej części obserwacji. Ponadto, nie ma wyraźnej zależności liniowej między zmiennymi y_t1 i y_t2 (jak wykazano w poprzednim punkcie raportu zmienne charakteryzują się niewielką korelacją), a wariancja punktów nie jest całkowicie stabilna w czasie, ponieważ punkty w środkowej części wykresu są gęściej rozłożone niż na jego brzegach. Na podstawie zebranych informacji zdecydowano się na odrzucenie tej metody imputacji.

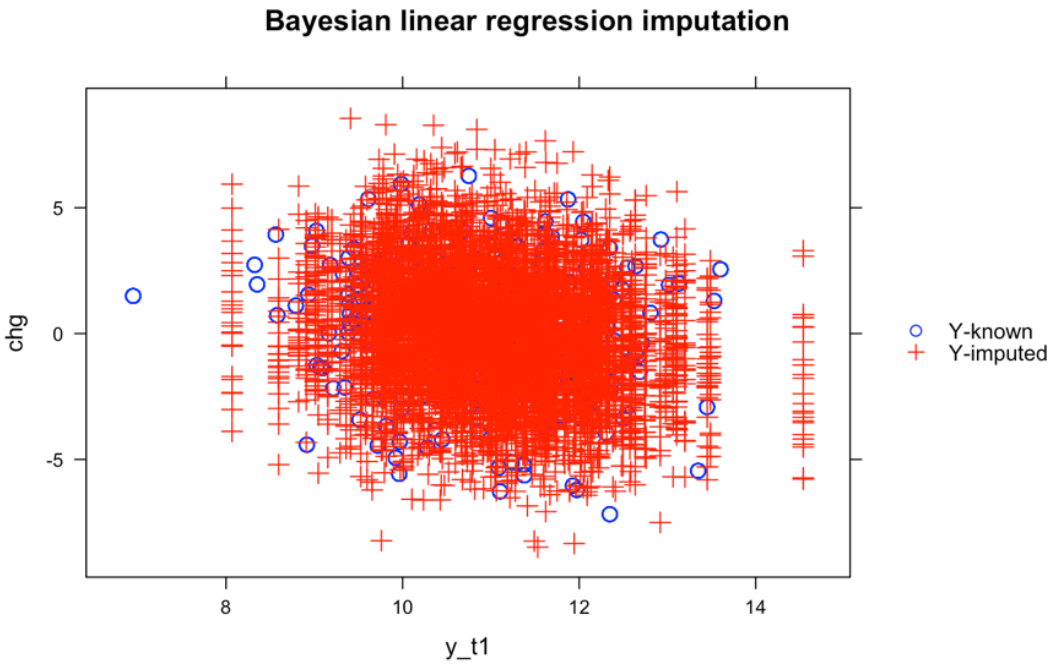


Wykres 10 Zależność zmiennej chg od zmiennej y_t1 dla y_t2 niebrakujących oraz zaimputowanych regresją liniową

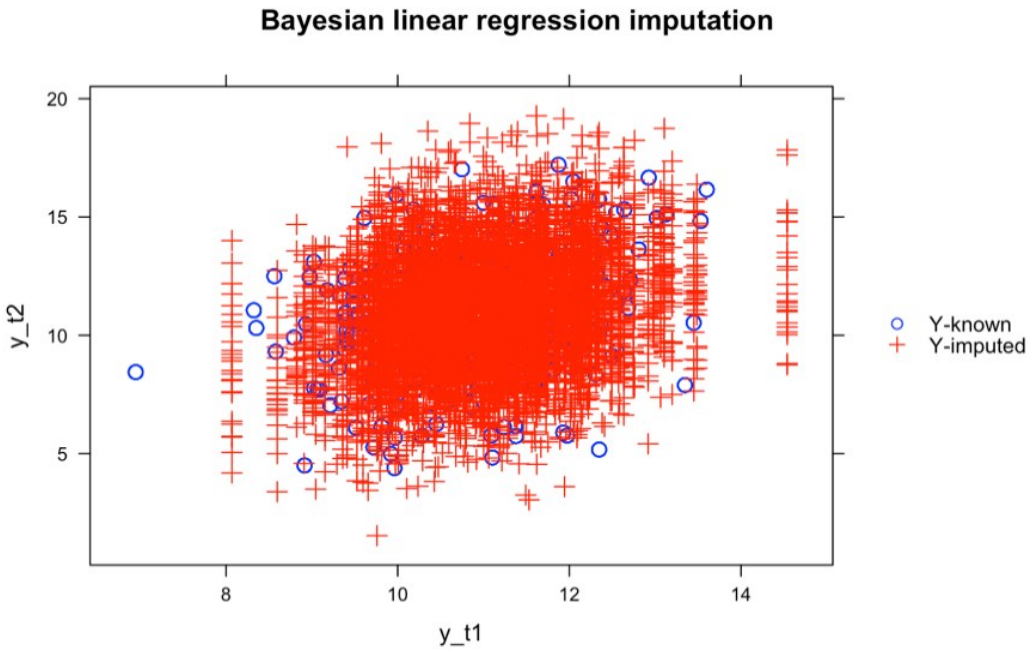


Wykres 11 Zależność zmiennej y_t2 od zmiennej y_t1 dla y_t2 niebrakujących oraz zaimputowanych regresją liniową

Wykresy dla bayesowskiej regresji liniowej przedstawiono na Wykresach 12 i 13. Można zauważyć, że imputowane zmienne są bardziej podobne do rzeczywistych wartości y_t2. Jednakże dane znajdują się poza naturalnym zakresem y_t2 znanych, co widać chociażby w dolnej części prawego wykresu w punkcie, który dla y_t2 kształtuje się wokół wartości 2, podczas gdy minimalna wartość y_t2 wynosiła 4,388 w szczególności w dolnej granicy. Z tego względu odrzucono tę metodę imputacji.

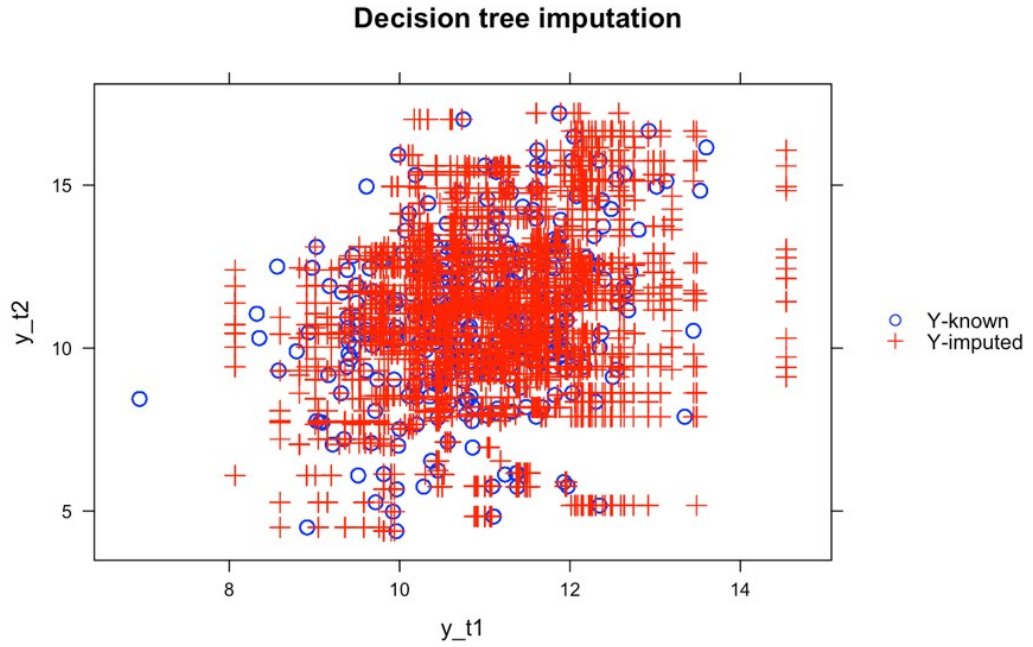


Wykres 12 Zależnosć zmiennej chg od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych w regresji liniowej bayesowskiej

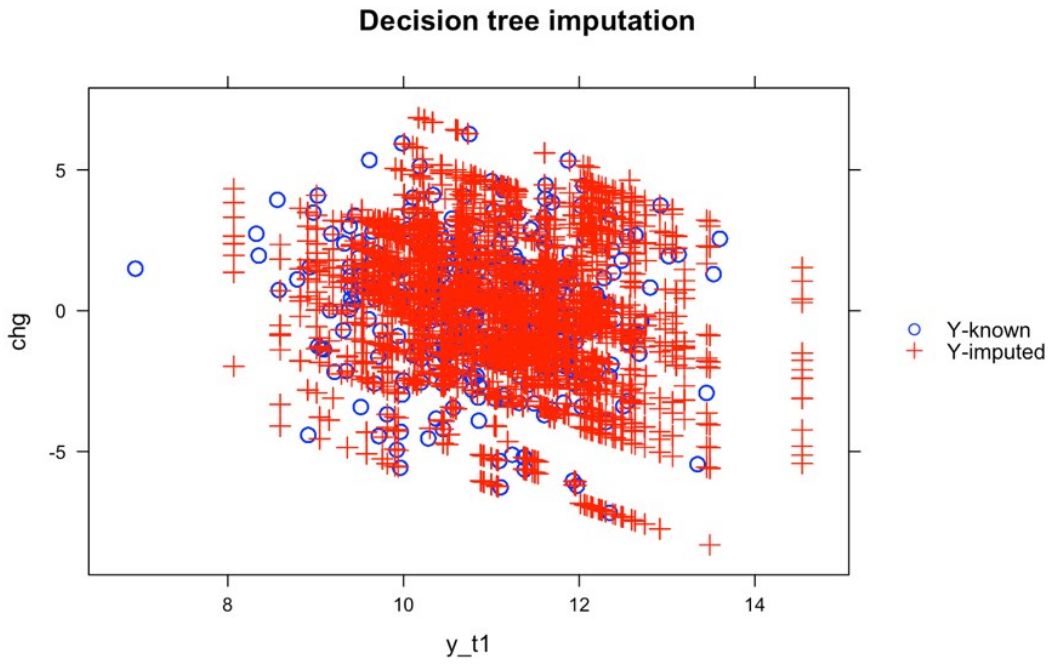


Wykres 13 Zależność zmiennej y_t2 od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych w regresji liniowej bayesowskiej

W trzecim przypadku wykorzystano metodę drzew decyzyjnych, której wynik zobrazowano na Wykresach 14 i 15. Dane imputowane są skoncentrowane niemal identycznie jak w przypadku znanych danych. Metoda ta została odrzucona, ponieważ technika ta jest zbyt eksploracyjna na potrzeby zadania.

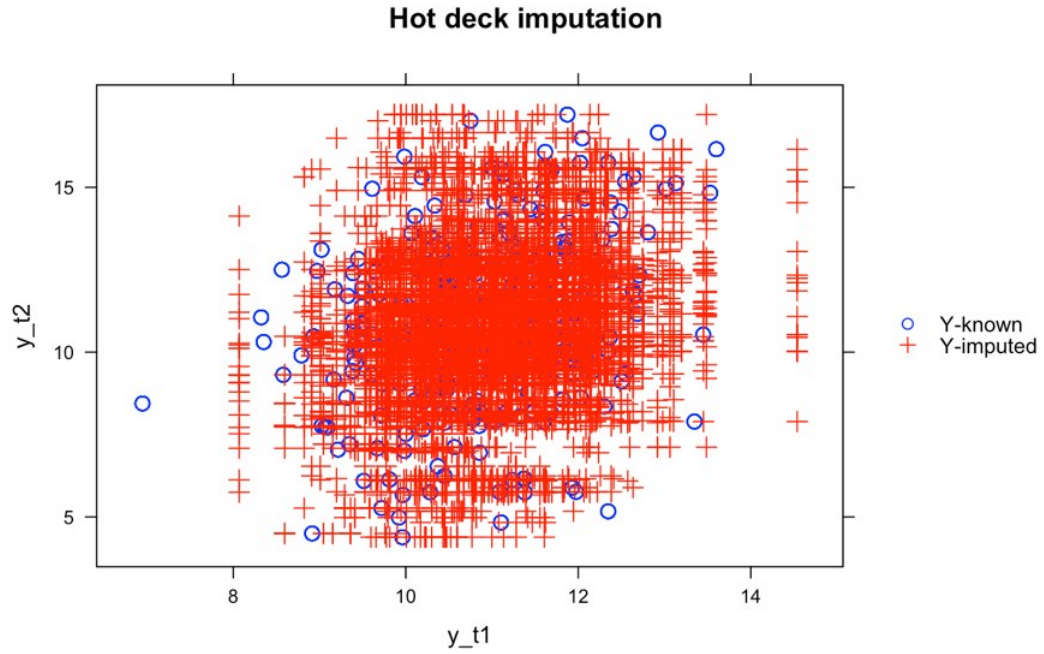


Wykres 14 Zależność zmiennej y_t2 od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych w drzewie decyzyjnym

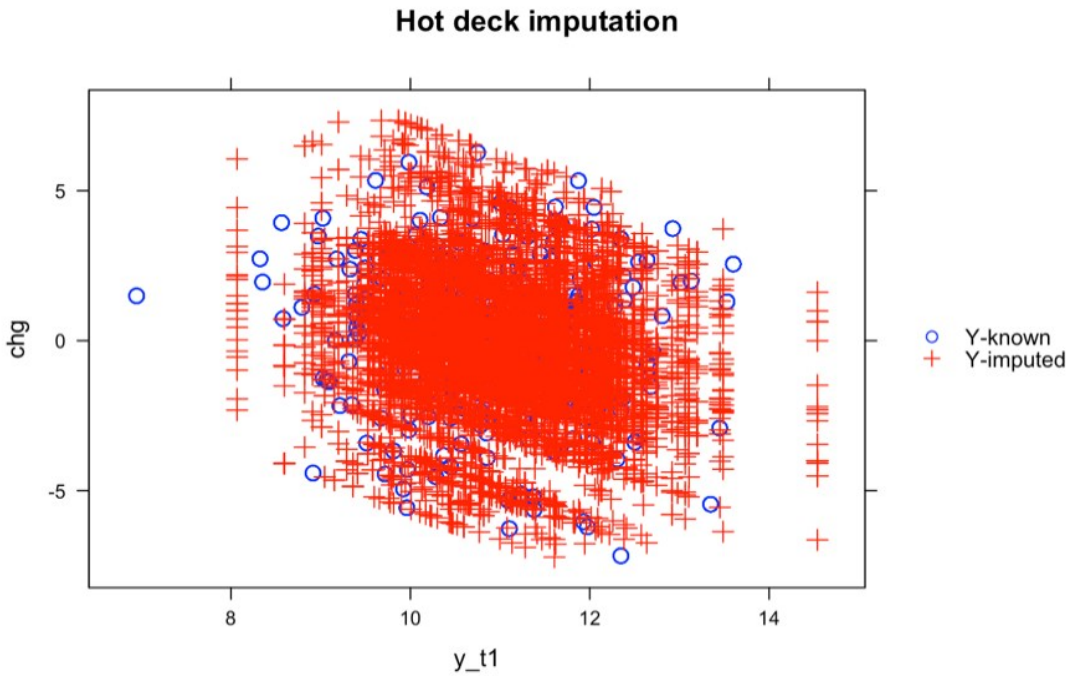


Wykres 15 Zależność zmiennej chg od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych w drzewie decyzyjnym

Ostatnią metodą imputacji była metoda „hot deck", której wynik pokazano na Wykresach 16 i 17. Dane mieszczą się w zakresie oryginalnych wartości danych i nie ma wielu wartości odstających. Po przeanalizowaniu wszystkich metod postanowiono wybrać metodę „*hot deck*" do dalszej analizy.



Wykres 16 Zależność zmiennej y_t2 od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych metodą hot deck



Wykres 17 Zależność zmiennej chg od zmiennej y_t1 dla y_t2 znanych oraz zaimputowanych metodą hot deck

Dla wybranej metody imputacji, tj. „*hot deck*", oszacowano model kowariancji MAR (*Missing at Random*). Dopasowano model regresji liniowej do imputowanego zbioru imp04, w którym zmienna objaśniana to różnica między zmiennymi y_t2 i y_t1, a predyktorami są zmienne y_t1 i trt. Funkcja *pool* używana jest do łączenia oszacowań parametrów z wielu imputacji. Otrzymane wyniki z funkcji *summary* przechowano w Tabeli 7.

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	3.5082527	1.3998853	2.506100	75.07207	1.437217e-02
2	y_t1	-0.3810163	0.1287070	-2.960338	70.28118	4.187181e-03
3	trt1	1.2566195	0.2441801	5.146282	83.75191	1.725873e-06

Tabela 7 Wyniki modelu

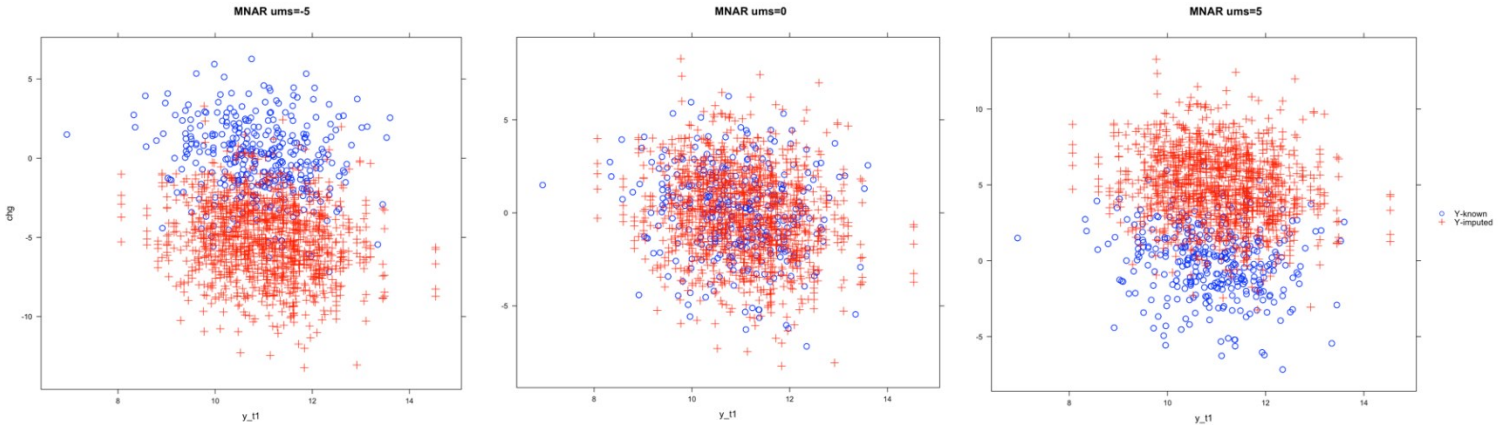
Na podstawie przeprowadzonej regresji liniowej stwierdzono, że oba badane parametry są istotne statystycznie przy ustalonym poziomie istotności 5%. Ponadto, zaobserwowano zbliżone wartości dla współczynników, porównując do sytuacji Missing Completely at Random.

3.3. Imputacja przy założeniach MNAR

W ostatniej części skupiono się na brakach danych typu MNAR (*Missing Not at Random*), gdzie rozkład braków danych m zależy od tych wartości. Mając na względzie test dotyczący założenia

MCAR, nie mamy podstaw twierdzić, że braki te są typu MNAR, jednak takie braki danych często są obecne w przypadku badań klinicznych prowadzonych w dwóch momentach.

Aby przeanalizować wyniki MNAR wykonano korektę oszacowań dla zestawu „przesunięć" wartości uzupełnionych (*tipping-point analysis*). W tym celu zbudowano funkcję *sensitivityanalysis*, w której przeprowadzono analizę wrażliwości przy użyciu imputacji wielokrotnej przez równania łańcuchowe (MICE) z metodą „mnar.norm" w zależności od wartości parametru „*MNAR ums*". Wewnątrz funkcji wykonano transformacje danych, wygenerowano wykresy oraz przeprowadzono regresję liniową w analogiczny sposób do działań opisanych podczas analizy przy użyciu różnych wariantów imputacji. Różnią się jedynie wartościami parametru „*ums*". Na potrzeby zadania wybrano przedział od -5 do 5, aby funkcja *sensitivityanalysis* iteracyjnie przechodziła po tych punktach. W taki sposób wygenerowano 11 wykresów zależności zmiennej chg od zmiennej y_t1 oraz 11 wykresów zależności zmiennej y_t2 od zmiennej y_t1 dla y znanych oraz zaimputowanych. W raporcie zamieszczono po 3 wykresy dla obu wersji z wartościami parametru „*ums*" równymi -5, 0 oraz 5 (Wykresy 18 i 19). Taki układ wykresów pozwala z łatwością zaobserwować „przesunięcia" wartości uzupełnionych (*tipping-point analysis*) wzdłuż osi rzędnych.

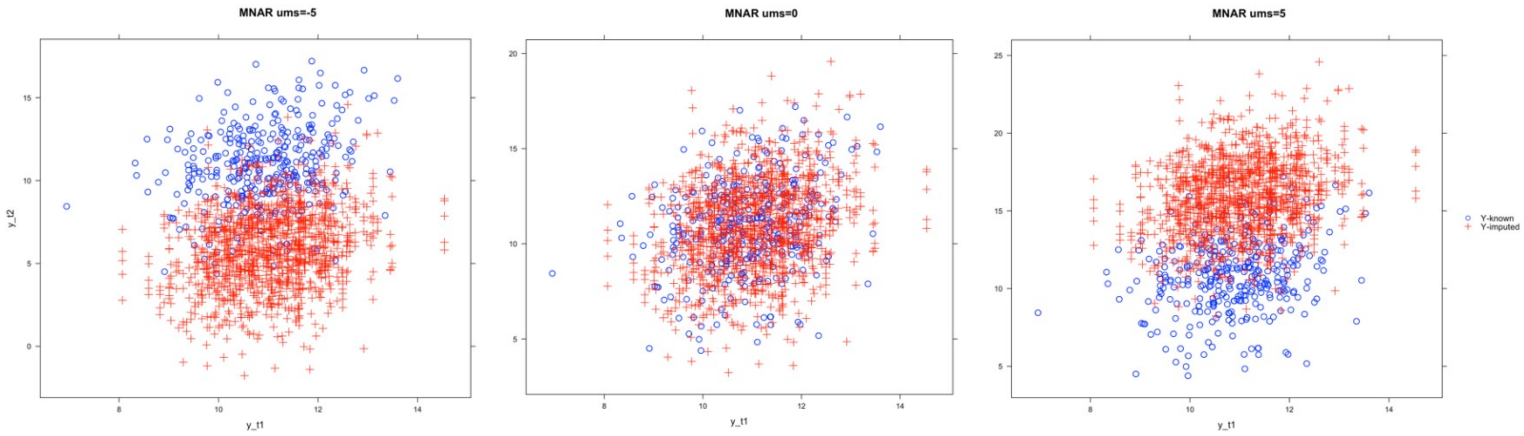


Wykres 18 Wykresy zależności zmiennych chg od zmiennej y_t1 dla y znanych oraz zaimputowanych w imputacji wielokrotnej dla ums = -5/0/5

Również w przypadku tego założenia możemy zaobserwować podobne wartości współczynników przy zmiennej trt dla modelu kowariancji. Współczynniki nie zmieniają się znacząco przy zmianie ums i wahają się od 1,19 do 1,03 (Tabela 8).

ums	-5	-4	-3	-2	-1	0	1	2	3	4	5
trt coeff	1,1986	1,1818	1,1650	1,1482	1,1314	1,1145	1,0977	1,0809	1,0641	1,0473	1,0305

Tabela 8 Współczynniki modelu kowariancji przy imputacji typu MNAR dla różnych wartości ums



Wykres 19 Wykresy zależności zmiennych y_t2 od zmiennej y_t1 dla y znanych oraz zaimputowanych w imputacji wielokrotnej dla ums = -5/0/5

4. Podsumowanie

Podsumowanie realizacji celów wyszczególnionych w projekcie:

1. Ustalenie wielkości średniej zmiany w zależności dla grup wystawionych i niewystawionych na działanie czynnika

Ustalenie wielkości średniej zmiany w podziale na grupy wykonano. W każdej z badanych sytuacji średnia zmiana charakteryzowała się istotną statystycznie dodatnią zależnością dla grupy wystawionych na działanie czynnika w stosunku do grupy kontrolnej. Wyniki wskazywały na względną stabilność i w zależności od wybranej metody wahały się od 1.25 (MAR) do 1.03 (MNAR). Największe zaufanie budzi jednak 1.18 (MCAR) z powodów przedstawionych dokładnie w powyższym raporcie.

2. Poznanie rozkładu i analizę graficzną badania

Rozkład i analiza graficzna danych zostały wykonane. Ich wyniki przedstawiono w punkcie 2. Analiza eksploracyjna danych nie wykazała nielosowości braków danych zmiennej y_t2 (tj. jedynej zmiennej z brakującymi obserwacjami w danym zbiorze).

3. Ocena zależności pomiędzy prawdopodobieństwem braku danych a pozostałymi cechami;

Ocena zależności pomiędzy prawdopodobieństwem braku danych a pozostałymi ocenami została przeprowadzona. Wyniki zaprezentowano w punkcie 2. Nie zaobserwowano zależności braku danych od wartości pozostałych zmiennych.

4. Oszacowanie modelu analizy kowariancji MCAR oraz MAR wraz z oceną zasadności stosowania założenia o nielosowym charakterze braków danych MNAR i wykonanie *tipping point analysis*;

Dokonano imputacji przy założeniu MCAR, MAR oraz MNAR i porównano wyniki współczynników przy zmiennej trt. Ze względu na test przeprowadzony w rozdziale 3.1 stwierdza się, że brak istotnych różnic w średnich predysponuje do imputacji danych zgodnie z założeniem MCAR. Warto również podkreślić, że we wszystkich przypadkach współczynniki były podobne co może oznaczać względną stabilność. Dla braków danych MAR wykorzystano różne warianty imputacji w celu ich przeanalizowania, m.in. regresję liniową, drzewo decyzyjne czy metodę „hot deck". Ostatnia z nich została wybrana do oszacowania modelu kowariancji. W przypadku braków danych MNAR dokonano korekty oszacowań dla zestawu „przesunięć" wartości uzupełnionych (*tipping-point analysis).* Współczynniki nie ulegają znaczącej zmianie przy modyfikacji parametru ums.

5. Podsumowanie i ocena wyników estymacji modelu właściwego w kontekście braków danych.

Na podstawie zarówno wyników analizy eksploracyjnej danych testu różnicy średnich oraz oszacowań modelu analizy kowariancji stwierdza się, że odpowiednim modelem braków danych w kontekście niniejszego zadania jest model MCAR.