**Zadanie I**

Zbiór: z1

Oszacuj średnią wartość Y na podstawie zbioru z1, korzystając z imputacji wielokrotnej, przy

założeniu, że dane pochodzą z wielowymiarowego rozkładu normalnego. **Skorzystaj z metody stochastycznej imputacji regresyjnej**. Przyjmij ziarno losowe równe 10. Liczbę imputacji ustal na 25. Pozostałe opcje powinny przyjmować wartości zgodnie z ustawieniami domyślnymi. Korzystając z otrzymanych wyników odpowiedz na pytania 10, 11 i 12.

Odp 1:

proc mi data=w seed=10 nimpute=25 out=outmi2 simple;

var X Y;

fcs regpmm ; /\* Czy to jest dobra metoda ?

run;

Odp 2:

**proc mi data=dane.z1 nimpute=25 seed=10 out=stochastic;**

**var x y;**

**monotone reg(y = x / details);**

**run;**

kod od prowadzacego:

\*Stochastic regression imputation;

\*http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug\_mi\_sect021.htm;

proc mi data=crop01 seed=1234 nimpute=1 out=stochastic;

var size wormy; \*NOTE: Order is important;

monotone reg(wormy = size / details);

run;

Odp 3:

proc mi data=e01 out=wynik1 nimpute=25 seed = 10;

var Y X;

mcmc;

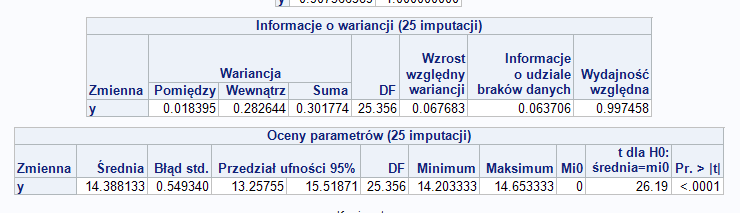
run;

**JAK MYŚLICIE Odp 1 czy Odp 2? Inne? A nie powinno być 3?**

**To która w końcu metoda jest dobra? - Odp.3, potwierdzam**

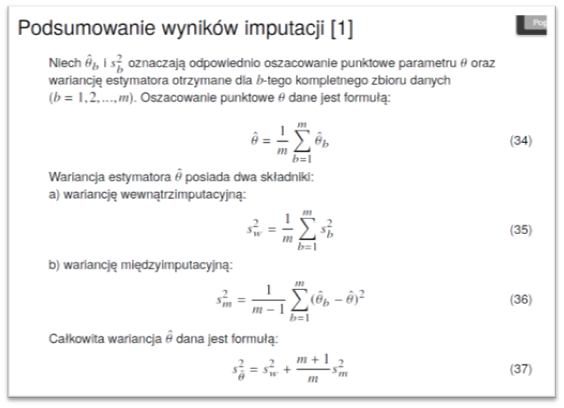
**a nie powinno byc 2? w 3 nie ma zdefiniowanej zadnej regresji.**

**MCMC to jest ta regresja**



10. Jaki jest udział wariancji międzyimputacyjnej w całkowitej wariancji estymatora średniej Y? =wariancja wewnątrz/suma

1. **5%-8%**
2. 11% i więcej
3. poniżej 5% ja bym zaznaczyła tą odpowiedz h
4. 8%-11% Skąd tę wartość się odczytuje ? i dlaczego akurat między 8-11 ? Dzielisz wariancje pomiędzy, przez sumę (z tabelki na górze) - **a to wtedy nie wychodzi 6%? Czy to nie jest 0,018/0,282?. To jest 0.018/0.30- czylo w sumie jest 5-8**



11. Podaj wartość oszacowania średniej Y po imputacji.

1. 16
2. 14
3. 18
4. 12

12. Odnosząc się do wyników imputacji wielokrotnej oraz oceny pomiędzy zmiennymi X i Y wskaż, które z poniższych zdań jest prawdziwe.

1. Przyjęcie założenia MCAR prowadziłoby do niedoszacowania zmiennej średniej Y
2. Przyjęcie założenia MCAR prowadziłoby do poprawnej oceny średniej Y
3. **Przyjęcie założenia MCAR prowadziłoby do przeszacowania zmiennej średniej Y**

**A czemu ta odpowiedź?**

**Odp1: Czy tu nie trzeba się porównać do średniej Y bez imputacji? Chyba tak**

**Czy jeśli zakładamy, że bez imputacji średnia wynosi 14.97, to jak mamy całkowitą losowość średnia po imputacji powinna być taka sama? A bo imputacji jest niższa wartość, więc model jest niedoszacowany?**

**Zadanie II**

Poniżej zaprezentowano wyniki oszacowań parametrów modelu z efektami stałymi dla danych wzdłużnych postaci:g

Y it =β 1 + β 2 \*trt+ β 3 \*t+ β 4 \*trt\*t +ε it

gdzie:

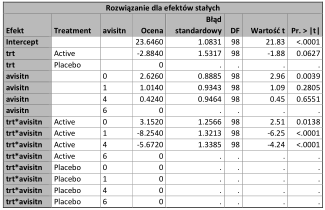
Y it – wyniki pomiaru Y dla i-tej jednostki w momencie t-tym (i=1,2,…,n; t=0,1,4,6),

trt – grupa, do której losowo przypisano jednostkę przed rozpoczęciem badania (1=Active,

2=Placebo),

t ≡ avisitn – moment pomiaru,

ε it – reszta losowa o zadanym rozkładzie prawdopodobieństwa

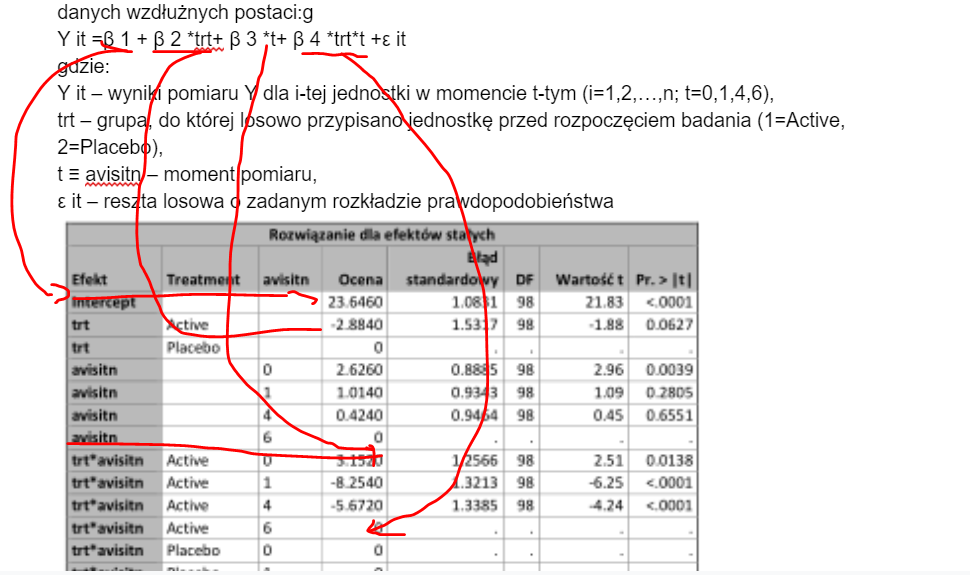


Na podstawie jakiej funkcji został wygenerowany ten output ?

On już był w zadaniu

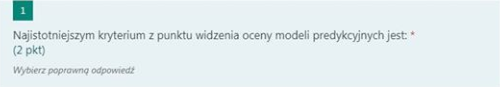
13. Korzystając z oszacowań parametrów modelu, podaj wartość średnią Y dla ostatniego pomiaru (t=6) w grupie poddanej działaniu czynnika “Active”.

**20,76 - skąd ten wynik? Intercept + trt active -> bo reszta jest 0, dlatego bierzesz tylko te parametry ( screen na dole)**

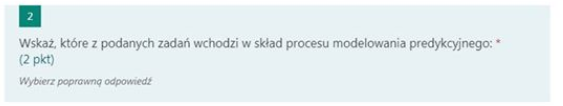
****

14. Czy średnia w grupie “Active” jest statystycznie istotnie różna od średniej w grupie “Placebo” w przypadku ostatniego pomiaru (t=6)?

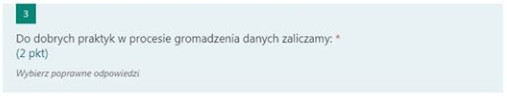
1. Tak ( a nie powinna być ta odpowiedz ? p-valeu 0.0627)u No właśnie p>0.05 czyli nieistotna
2. **Nie**
3. Nie można stwierdzić na podstawie podanych wyników.



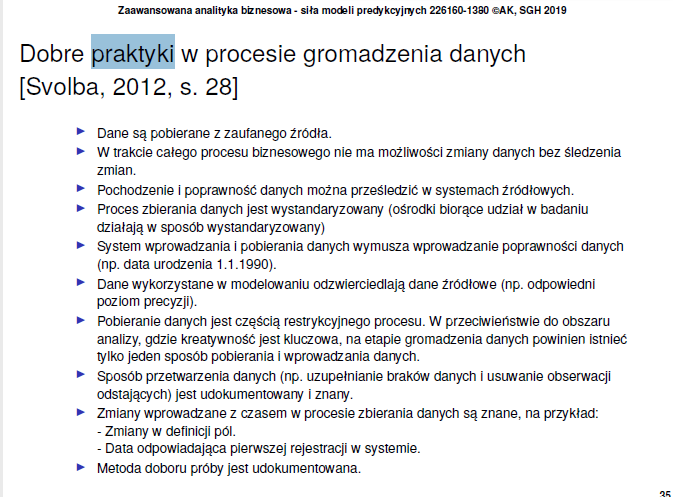
1. interpretowalność oszacowań estymowanych parametrów
2. **zdolność do formułowania dokładnej prognozy**
3. własności estymatorów
4. prostota formułowanych modeli

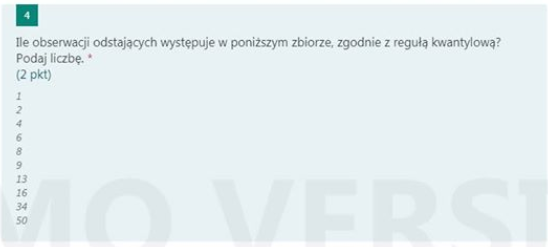
4

1. podział próby
2. ocena jakości danych
3. imputacja danych
4. **wszystkie podane**



1. dokumentowanie przyczyny czas i autora zmiany bazy danych
2. wymuszanie wprowadzania danych z określonych okresów
3. automatyczne uzupełnianie braków danych
4. **wszystkie podane**

****



Odp: 34 i 50 (wartości odstające powyżej IQR)

- mi wyszedł koniec przedziału 31,4. A jak to liczyłeś?

Na piechotę: Q1 = (2 + 4) / 2; Q3 = (13+16) / 2; IQR = 11.5; 1.5 \* IQR = 17,25

Q3 + IQR \*1,5 = 31, 75

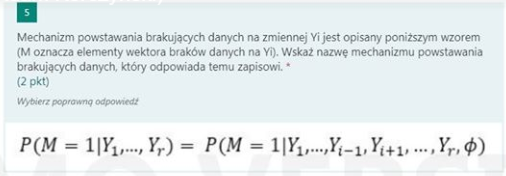
Q1 - IQR \* 1,5 = -14,25

Mi tak samo, używam tego koda

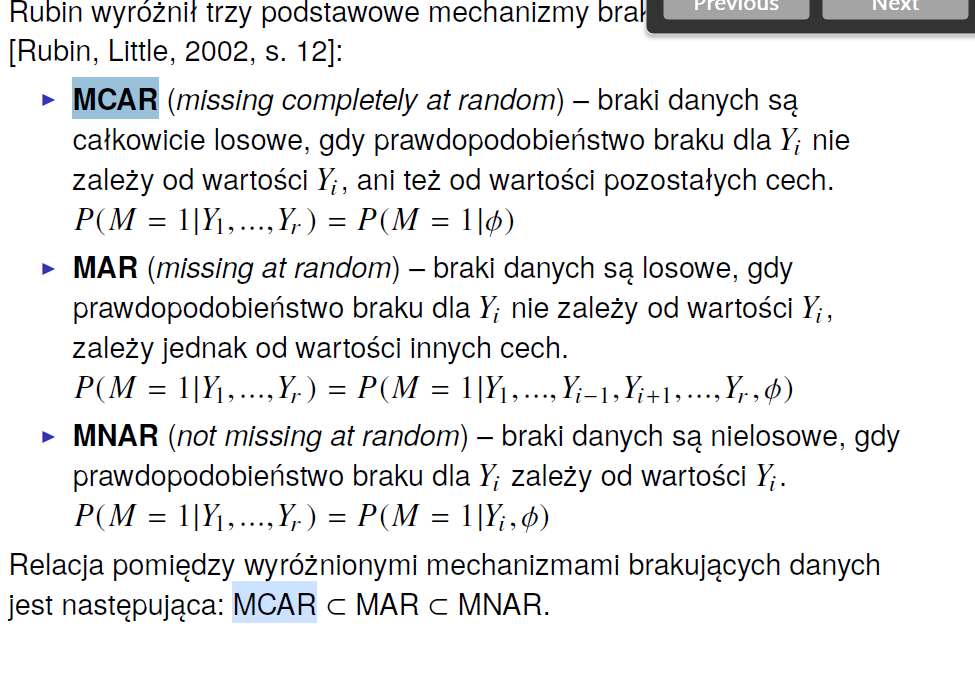
w <- c(1, 2, 4, 6, 8, 9, 13, 16,34,50)

IQR <- IQR(w)

start <- quantile(w, 0.25) - 1.5 \* IQR

stop <- quantile(w, 0.75) + 1.5 \* IQR

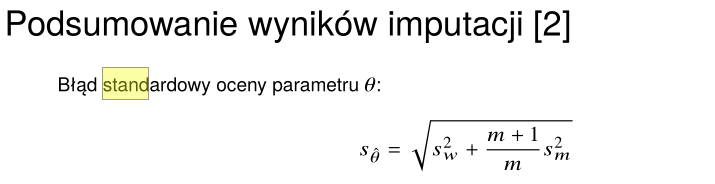
1. Braki danych całkowicie losowe MCAR
2. Braki danych losowe MAR
3. Braki danych nielosowe MNAR



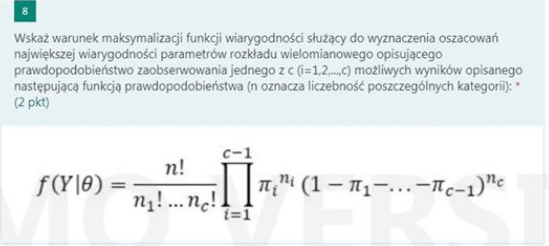


1. **Służy do oceny zróżnicowania wynikającego z losowego doboru próby**
2. Służy do oceny zróżnicowania oszacowań na przestrzeni imputacji
3. a i b
4. żadne z powyższych





1. 
2. 
3. 
4. 

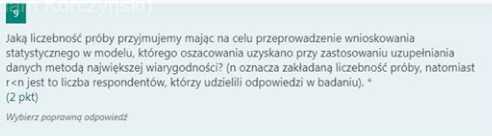


a)

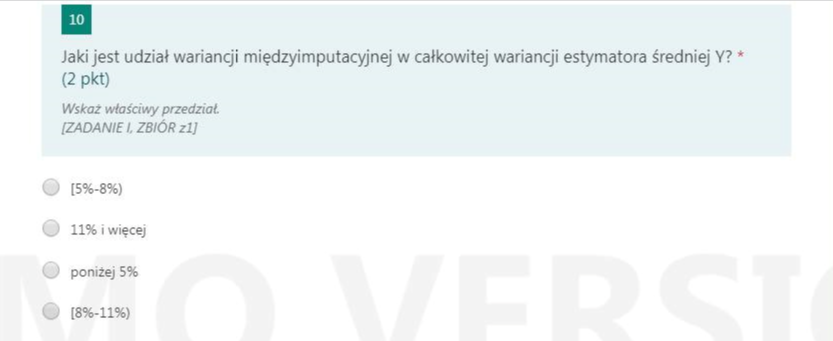
b) 

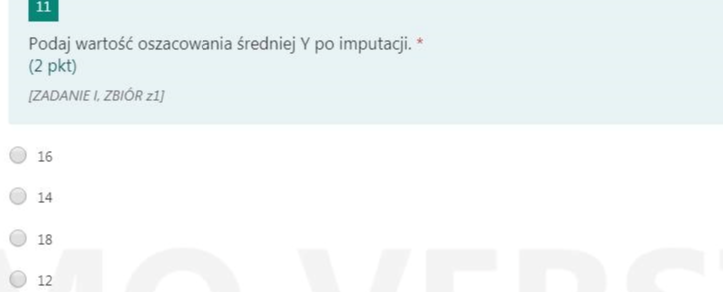
c) 

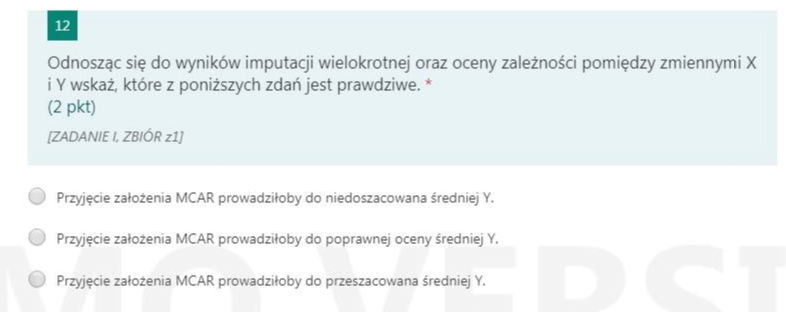
d) 



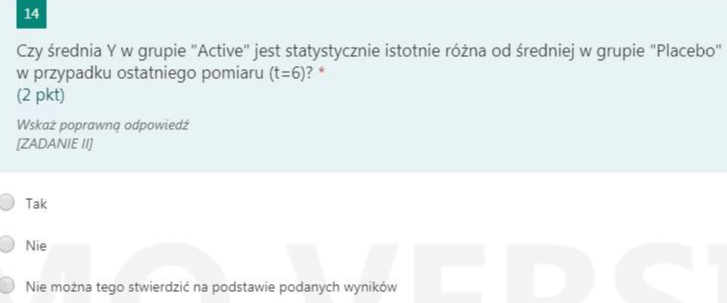
1. n-r
2. n
3. żadna z odpowiedzi
4. r

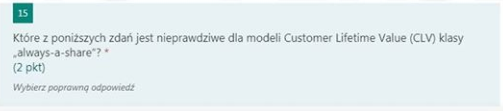




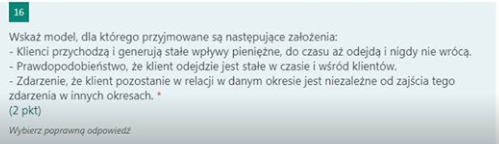




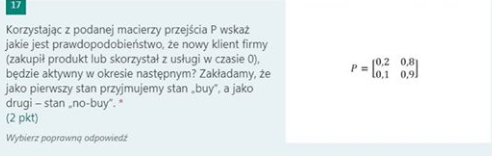




1. Przykładem jest uogólniony model retencji
2. Stosuje się w przypadku, gdy analiza powinna uwzględniać klientów, którzy ponownie się uaktywnili
3. Nie zakłada, że brak aktywności ze strony klienta oznacza jego stałe odejście
4. Przykładem jest model migracji typu recency -frequency



1. model migracji
2. Prosty model retencji
3. Uogólniony model retencji
4. Model renty klienckiej



1. 0.9
2. 0.2
3. 0.8
4. 0.1

**Zadanie III**

Zbiór: churn\_dataset

Opis zmiennych w zbiorze:

id – zmienna identyfikująca klientów

account\_length – długość trwania umowy w tygodniach

churn – 0 gdy klient nie odszedł, 1 – gdy klient odszedł

international\_plan – ‘yes’ – klient wykupił pakiet międzynarodowy, ‘no’ – klient nie wykupił

pakietu międzynarodowego

Korzystając ze zbioru danych churn\_dataset oszacuj uogólnione modele retencji:

- bez stratyfikacji, oraz

- ze stratyfikacją po zmiennej international\_plan. Skorzystaj z estymatora Kaplana-Meiera. Na podstawie wyników oszacuj CLV przy założeniu tygodniowej opłaty w wysokości 30 j.p. oraz stopy dyskontowej 1%.

Korzystając z otrzymanych wyników odpowiedz na pytania 18, 19 i 20.

A jaki jest kod który to robi?

Odp1: z kodów do modeli retencji (inny zbiór)

Próbowałem to tak skopiować, ale nie wiedziałem jak te zmienne podmienić

/\*Oszacowanie metoda Kaplana-Meiera\*/

proc lifetest data=dane.churn\_dataset outsurv=out1;

time account\_length\*churn(0);

run;

/\*Oszacowanie metoda Kaplana-Meiera ze stratyfikacja\*/

proc lifetest data=dane.churn\_dataset outsurv=out2;

strata international\_plan;

time account\_length\*churn(0);

run;

/\*Oszacowanie CLV\*/

data wyn2;

set out2;

by account\_length;

if account\_length>0 AND first.account\_length;

clv = survival\*30/(1.01\*\*(account\_length-1));

run;

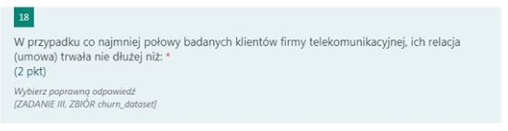
proc print data=wyn2 noobs;

var account\_length survival clv;

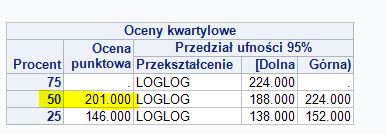
sum clv;

format clv dollar8.2;

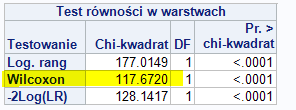
run;



1. 146 tygodni
2. 181 tygodni
3. 201 tygodni
4. 243 tygodnie







1. 117.67
2. 128.14
3. 177.01
4. żadna z odpowiedz



1. Osoby z wykupionym pakietem międzynarodowym mają średnio dłuższy czas pozostawania w relacji z firmą niż osoby, które tego pakietu nie wykupiły.
2. CLV pojedynczego klienta dla pierwszych 12 tygodni trwania umowy wynosi w przybliżeniu 340.82 j.p. A nie będzie to? Trzeba zrobić cumulative, przy 30 j.p per okres powinno być to raczej więcej niż 26 j.p. jak w odpowiedzi d)

Też mi się wydaje, że jest to bardziej logiczne. Czy ktoś mógłby podpowiedzieć w jaki sposób oblicza się to cumulative ? Musisz wziąć sumę wszystkich płatności od 1 do 12 włącznie

1. żadna z odpowiedzi
2. CLV pojedynczego klienta dla pierwszych 12 tygodni trwania umowy wynosi w przybliżeniu 26.87 j.p.

Pytania z Menti:

Do klasy modeli typu gone for good zaliczamy:

a) Model migracji

**b)** **Prosty model retencji**

c) Podejście data mining do wartość w czasie

d) Żadna nie jest prawidlowa

W modelu k- średnich wariancja błędu losowego jest:

**a)** **Wspólna dla wszystkich klastrów i zmiennych**

b) Zmienna względem klastrów

c) Zmienna względem zarówno zmiennych jak i klastrów

d) Zmienna względem zmiennych

Przykładem modelu deterministycznego jest:

**a)** **Model renty klienckiej**

b) Prosty model retencji

c) Uogólniony model retencji

d) Żaden z powyższych

Oszacowania CLV modelem prostym w porównaniu do modelu uogólnionego są:

a) Zawsze wyższe

b) Zawsze niższe

c) Losowo raz wyższe, raz niższe

**d)** **Zależy np. od kształtu funkcji dożycia**

Do założeń uogólnionego modelu retencji nie należy:

**a)** **Niezmienność stopy retencji w czasie**

b) Zmienność stopy retencji oraz przepływów finansowych

c) Zależność przepływów finansowych od momentu rezygnacji klienta z umowy

d) Brak możliwości powrotu klienta po odejściu

Czy odpowiedź d) też nie jest poprawna?