1. **Wyjaśnij czym jest modelowanie predykcyjne? Podaj podstawowe kryterium oceny modeli predykcyjnych. [2p]**

Modelowanie predykcyjne to proces rozwijania narzędzia matematycznego lub modelu, generującego dokładną prognozę wykorzystując odpowiednie dane, wiedzę ekspercką oraz wykorzystując odpowiednie narzędzia statystyczne .

Ocenę modelu predykcyjnego dokonujemy poprzez porównania wyników zbioru walidacyjnego z wynikiem otrzymanym na zbiorze treningowym. Zaczynamy od modeli które potrafią odzwierciedlić złożone zależności. Jeśli wyniki są podobne między dostępnymi modelami, wybieram model który jest lepiej zrozumiały lub mniej skomplikowany co ułatwią interpretacje wyników.

Slajd 8, 27

1. **Opisz proces walidacji modelu predykcyjnego. [2p]**

W statystyce walidacja modelu polega na potwierdzeniu, że wyniki modelu statystycznego są dopuszczalne w odniesieniu do rzeczywistego procesu generowania danych.

Przykład techniki próbkowania wtórnego: k-krotna walidacja krzyżowa

1. Podział próby na k zbiorów o mniej więcej równej wielkości

2. Dopasowanie modelu z wykorzystaniem wszystkich obserwacji spoza pierwszego zbioru

3. Predykcja na podstawie zbioru pierwszego w celu oceny efektywności modelu

4. Zwrócenie pierwszego zbioru do zbioru uczącego i powtórzenie punktów 1-3 z usunięciem drugiego zbioru

5. Podsumowanie k oszacowań dla każdego zbioru

Metody próbkowania wtórnego pozwalają uzyskać dość dobre prognozy tego, jak dobrze model sprawdzi się na przyszłych próbach danych.

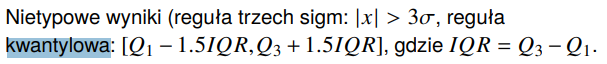
1. **Wyjaśnij czym są metadane i jakie mają znaczenie w kontekście oceny jakości danych? [2p]**

Metadane informują o jakości zbioru, stanowią informacje opisujące zbiór danych. Metadane zawierają informacje o zawartości danych, ich datach, a więc aktualności.

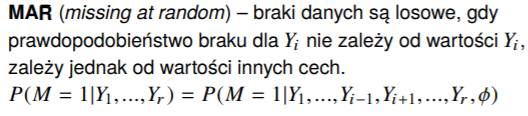
Dane o danych – ustrukturalizowane informacje stosowane do opisu zasobów informacji lub obiektów informacji, dostarczają szczegółowych danych, dotyczących atrybutów zasobów lub obiektów informacji, w celu ułatwienia ich znalezienia, identyfikacji, a także zarządzania tymi zasobami.

1. **Nazwij oraz podaj wzory opisujące dwie reguły służące do wykrywania obserwacji odstających [2p]**

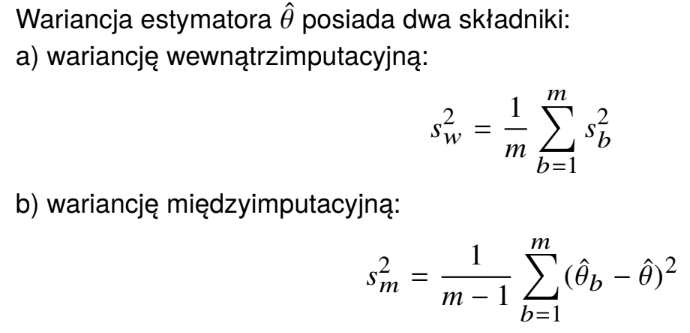
* reguła trzech sigm: |x| > 3**σ**,
* reguła kwantylowa: [Q1 − 1.5IQR,Q3 + 1.5IQR], gdzie IQR = Q3 −Q1.



1. **Mechanizm powstawania brakujących danych na zmiennej Yi jest opisany poniżej wzorem (M oznacza elementy wektora braków danych na Yi). Wskaż nazwę mechanizmu powstania brakujących danych, który odpowiada temu zapisowi. Jakie jest praktyczne znaczenie takiego założenia? [2p]**

****

1. **Wymień i wyjaśnij dwa składniki wariancji w metodzie imputacji wielokrotnej. [2p]**

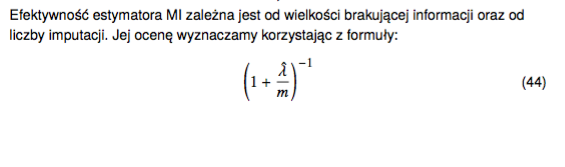
****

Wyjaśnienie:

Wariancja wewnątrzimputacyjna odzwierciedla zróżnicowanie wynikające z losowego doboru próby, zaś wariancja międzyimputacyjna pozwala ocenić zróżnicowanie pomiędzy oszacowaniami otrzymanymi dla poszczególnych zbiorów danych.

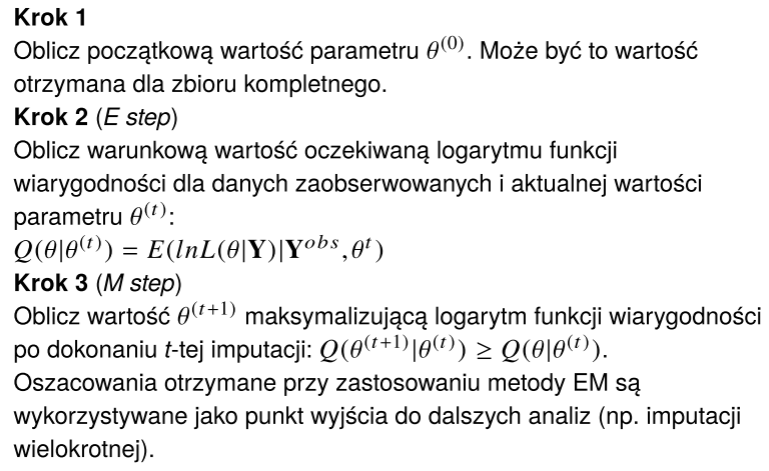
1. **Podaj statystykę, która służy do wyboru liczby imputacji w metodzie imputacji wielokrotnej. Podaj zakres zmienności oraz interpretację tej statystyki. [2p]**

Może chodzi o efektywność estymatora?

****

Wydaje mi się, że zakres zmienności dla niej powinien być od 0 do 1 i staramy się znaleźć takie m, które będzie maksymalizowało tą statystykę.

1. **Podaj nazwę i opisz podstawowe kroki algorytmu służącego do wyznaczania oszacowań największej wiarygodności dla danych niekompletnych. [3p]**

Algorytm EM Expectation-Maximization

1. **Jaką liczebność próby przyjmujemy mając na celu przeprowadzenie wnioskowania statystycznego modelu, którego oszacowania uzyskano przy zastosowaniu uzupełniania danych metodą największej wiarygodności? (n oznacza zakładaną liczebność próby, natomiast r<n jest to liczba respondentów, którzy udzielili odpowiedzi w badaniu). Wybierz poprawną odpowiedź [1p]**
2. ***n***
3. **r**
4. **n-r**
5. **żadna z powyższych odpowiedzi nie jest prawidłowa**
6. 

**10.1 Jaki jest udział wariancji międzyimputacyjnej w całkowitej wariancji estymatora? [1p]**

0.018271/0.160358=0.113938

**10.2 Podaj wartość oszacowania średniej Y po imputacji. [1p]**

15.022702

**10.3 Zweryfikuj hipotezę głoszącą, że średnia Y wynosi 10 wobec hipotezy alternatywnej, że jest różna od 10. Podaj odpowiednie miary uzasadniające decyzję weryfikacyjną. [1p]**

Ho: Mean=Mu0 gdzie Mu0=10

H1: Mean ≠ Mu0 gdzie Mu0=10

Przyjmując poziom istotności α=0,05 odrzucamy hipotezę zerową (ponieważ (Pr>|t|) <α) na rzecz hipotezy alternatywnej, tzn. Średnia Y jest różna od 10.

**10.5**

proc mi data=OUT.Z2 seed=10 mu0=10 ;

Monotone reg(y);

var x y?;

run;

1. **Na czym polega regresja kwantylowa? Podając opis metody wskaż najistotniejsze różnice pomiędzy modelem regresji kwantylowej, a modelem regresji liniowej. [2p]**

Regresja kwantylowa pozwala estymować efekty odpowiadające kwantylom rozkładu zmiennej losowej np. kwartyle, kwintyle, medianę. Jest to model zbliżony do modelu regresji liniowej z tym tylko, że regresja kwantylowej dostarcza serię oszacowań opisujących proces, a nie tylko ocenę tendencji centralnej.

Interpretacja jest podobna do regresji liniowej, w regresji kwantylowej otrzymujemy jednak rodzinę linii co pozwala skupić się na wybranym przedziale rozkładu badanej zmiennej.

Jako główne zalety modelu regresji kwantylowej wymienić możem:

\*Odporność na obserwacje odstające.

\* Wgląd w wybrane fragmenty warunkowego rozkładu zmiennej zależnej



**12**.

W rozkładzie warunkowym kwantyla 0.75 błąd prognozy zmiennej InCost rośnie o około 0,95% natomiast błąd pierwszej kategorii przedsiębiorstwa zmniejsza błąd prognozy o około 0,15% **wtf czemu to sa procenty?**

**A czy nie powinniśmy tego przypadkiem interpretować jak regresji liniowej (tylko dla 3ciego kwantyla ofc)? Wydaje mi się, że w slajdach była mowa o błędzie prognozy, bo to błąd prognozy był zmienną objaśnianą. Czyli: Jeżeli łączne koszty pracy rosną o 1%, to wielkość sprzedaży dla 3ciego kwartyla wzrośnie średnio o 0,95 % ceteris paribus.**

**Czy ktoś wie, jak się to interpretuje dla pierwszej kategorii przedsiębiorstwa?**

**13. Które z założeń jest nieprawdziwe względem modeli estymujących CLV klasy “always a share”[1p]**

1. **nie zakładają, że brak aktywności ze strony klienta oznacza jego stałe odejście**
2. **stosuje się w przypadkach, gdy analiza powinna uwzględniać klientów którzy się uaktywnili**
3. **przykładem jest uogólniony model retencji**
4. **przykładem jest model migracji typu ‘recency-frequency’**

**14. Jakie jest prawdopodobieństwo, że nowy klient firmy (zakupił produkt lub skorzystał z usługi firmy w czasie 0), będzie aktywny w okresie następnym w przypadku macierzy przejścia o postaci podanej poniżej? Zakładamy, że pierwszy stan to “buy”, a drugi - stan “no-buy”. [2p]**

**P= (0,2 0,8)**

**(0,1 0,9)**

1. **0,2**
2. **0,8**
3. **0,3**
4. **0,9**

**15. Zbudowano dwa modele retencji estymujące wartość CLV: model prosty i uogólniony. Model prosty dał oszacowanie bardziej optymistyczne (wyższe) niż model uogólniony. Wskaż z czego może wynikać ta różnica. Czy oszacowanie uzyskane modelem prostym z reguły będą wyższe? [2p]**

Chodzi o to w jaki sposób oba modele traktują odejście klienta w czasie. Model prosty zakłada ze klient jest z nami przez cały okres i odchodzi dopiero na jego końcu. Model uogólniony bierze pod uwagę odejście klienta w trakcie trwania okresu.

Ze względu na to zarówno jeden jak i drugi model może przeszacować lub niedoszacowac.

**16. Wśród omawianych typów modeli retencji i migracji podaj przykład jednego deterministycznego i jednego probabilistycznego modelu estymującego wartość CVL. [2p]**

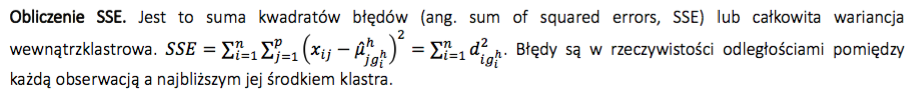
1. **model deterministyczny: model renty klienckiej**
2. **model probabilistyczny: uogólniony model retencji**

**17. Zbudowano model segmentacji z wykorzystaniem PROC FASTCLUS. W wynikach uzyskano poniższe statystyki. Wskaż, która z wartości odpowiada błędowi modelu. Uzasadnij swój wybór [2p]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Statistics for Variable** | | | | |
| **Variable** | **Total STD** | **Within STD** | **R-Square** | **RSQ/(1-RSQ)** |
| **x** | 2.06854 | 0.89098 | 0.823609 | 4.669219 |
| **y** | 1.02113 | 1.00352 | 0.039039 | 0.040683 |
| **OVERALL** | 1.63119 | 0.93959 | 0.669891 | 2.029303 |

Overall Within STD

Wydaje mi się, że to wynika z poniższego fragmentu na temat K-Means ale nie jestem pewna.



**18. Zapisz kod programu, który generuje nieparametryczny model retencji Kaplana-Meiera, a wyestymowane wartości funkcji przeżycia zapisze do zbioru wynikowego o nazwie outsurv w bibliotece WORK. Zbiór danych o nazwie retention znajduje się w bibliotece WORK, zmienne oznaczają długość czasu trwania relacji nosi to rel\_lenght, a zmienna cenzurująca to censor, gdzie 1 oznacza brak ocenzurowania, a 0 - obserwacje ocenzurowane. Dodatkowo w kodzie uwzględnij test na istotność wpływu zmiennej jakościowej *source* (źródło pozyskania klienta) na wartości przeżycia [1p]**

**proc lifetest data=work.retention outsurv=outsurv;**

**time rel\_lenght\*censor(0);**

**freq count;**

**strata source;**

**run;**

1. **Po wykonaniu testu na istotność wpływu zmiennej jakościowej *source* uzyskano następujące wyniki. Dokonaj interpretacji poniższych oszacowań, wskazując hipotezę zerową i alternatywną [1p]**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test of Equality over Strata** | | | |
| **Test** | **Chi-Square** | **DF** | **Pr >**  **Chi-Square** |
| **Log-Rank** | 25.4037 | 3 | <.0001 |
| **Wilcoxon** | 19.4331 | 3 | 0.0002 |
| **-2Log(LR)** | 33.9343 | 3 | <.0001 |

Ho: Wartość przeżycia jest taka sama dla wszystkich grup, niezależnie od źródła pozyskania klienta (od wartości zmiennej source).

H1: Wartość przeżycia jest różna w zależności od źródła pozyskania klienta.

Przyjmując poziom istotności α=0,05 odrzucamy hipotezę zerową (ponieważ wartości Pr>Chi-Square dla wszystkich testów <α) na rzecz hipotezy alternatywnej, tzn. Wartość przeżycia jest różna w zależności od źródła pozyskania klienta.