Sprawozdanie 1 z przedmiotu Statystyczna Analiza Danych

Tomasz Biegus

30 marca 2019

1 Analiza

1.1 Sprawdzenie danych

Przed przystąpieniem do analizy należy sprawdzić czy w danych nie występują wartości, które z pewnością są wynikiem błędu. Zwykle pewność można mieć kiedy wartości nie zgadzają się z sensem fizycznym badanej zmiennej, dobrym przykładem jest ujemna wartość wieku. Ponadto w danych mogą występować braki, które należy odpowienio obsłużyć bądź przez usunięcie obserwacji zawierających brakujące wartości, bądź stosując jedną z metod uzupełniania braków.

W badanym zbiorze danych HOMA występuje 10 zmiennych:

- Płeć,
- WIEK,
- BMI,
- WHR,
- TROJGLICERYDY,
- HOMA,
- LDL.HDL,
- FAT.ALL.P,
- FAT.A.P,
- FAT.G.P.

Zgodnie z opisem w treści zadania, w badaniu brało udział 194 kobiety i 84 mężczyzn w wieku od 20 do 40 lat. Sprawdzono, że wszystkie obserwacje znajdują się w tym przedziale z wyjątkiem jednej, dla której wiek wynosił -34 lata. Najprawdopodobniej wartość ta jest skutkiem omyłkowego pojawienia się

znaku "minus", zatem obserwacji tej nie usunięto, a jedynie zmieniono wartość wieku na dodatnią.

Sprawdzono, że wszystkie wartości dotyczące poziomu tłuszczu w organizmie (całkowitego, gynoidalnego i androidalnego) zawierają się w sensownym przedziale [0, 100].

Zgodnie z treścią zadania, w badaniu brać powinny jedynie osoby o wskaźniku BMI zawierającym się w przedziale [18, 25]. Znaleziono w zbiorze 11 obserwacji które nie spełniały tego warunku. Usunięto je ze zbioru danych.

Sprawdzono, że wartości WHR, stężenia trójglicerydów i stosunku LDL/HDL wszystkie są nieujemne tak jak to być musi w rzeczywistości.

Po powyższych operacjach pozostało 265 obserwacji, z czego 259 kompletnych (bez braków). Liczba 6 wierszy z brakami stanowi $\frac{6}{259} \approx 2.3\%$ całego zbioru danych. Przy takiej ilości uzasadnione jest po prostu odrzucenie wierszy z brakującymi danymi zamiast podejmowania prób ich uzupełniania. W związku z powyższym usunięto 6 wierszy z brakami ze zbioru danych.

1.2 Wstępna analiza danych

1.3 Wylosowanie danych testowych

Ustawiono wartość ziarna i wylosowano 10 obserwacji, które nie uczestniczyły w dopasowywaniu modelu. Z danych wykorzystywanych w dalszym toku prac usunięto te 10 obserwacji otrzymując zbiór danych o liczności 249.

1.4 Zbudowanie modelu regresji

W oparciu o wszystkie numeryczne zmienne objaśniające zbudowano model regresji. Zmienną objaśnianą jest HOMA, zmiennymi objaśniającymi są:

1. WIEK,
2. BMI,
3. WHR,
4. TROJGLICERYDY,
5. LDL.HDL,
6. FAT.ALL.P,
7. FAT.A.P,
8. FAT.G.P.

Czyli wszystkie z wyjątkiem płci, która jest zmienną kategoryczną a nie numeryczną.

1.5 Analiza reszt dla pełnego modelu

1.6 Identyfikacja i usunięcie obserwacji odstających i wpływowych

1.7 Wybór modelu

1.8 Interpretacja współczynników wybranego modelu

1.9 ponowna analiza reszt dla wybranego modelu

2 Wnioski