Analiza Dużych Zbiorów Danych

Lista nr.1 - Wielokrotne testowanie

1. Wygeneruj macierz planu $X_{1000\times 950}$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Następnie wygeneruj wektor odpowiedzi zgodnie z modelem

$$Y = X\beta + \epsilon \ ,$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$ a $\epsilon \sim N(0, I)$.

Wykonaj następujące analizy w oparciu o modele wykorzystujące

i) pierwszych 5 zmiennych

ii) pierwszych 10 zmiennych

iii) pierwszych 20 zmiennych

iv) pierwszych 100 zmiennych

v) pierwszych 500 zmiennych

vi) wszystkie 950 zmiennych.

Dla każdego z powyższych modeli wyznacz estymator najmniejszych kwadratów dla wektora β i wykonaj testy istotności jego elementów. Porównaj jak się zmienia odchylenie standarowe estymatora β_1 i szerokość 95% przedziału ufności dla tego parametru w miarę tego jak rośnie rozważany model. Porównaj liczbę prawdziwych i fałszywych odkryć dla różnych modeli. Porównaj z liczbą fałszywych i prawdziwych odkryć po zastosowaniu korekt Bonferroniego i Benjaminiego Hochberga na wielokrotne testowanie.

2. Powtórz powyższe doświadczenie 1000 razy i dla różnych modeli wyznacz

a) średnią wariancję estymatora β_1 - porównaj z wartością teoretyczną (patrz odwrotny rozkład Wisharta)

b) średnią szerokość 95% przedziału ufności dla β_1 - porównaj z teoretycznym oszacowaniem

c) średnią liczbę prawdziwych i fałszywych odkryć oraz estymatory FWER i FDR dla procedur testowania bez korekty oraz z korektą Bonferoniego i BH. Dla procedur bez korekty i z korektą Bonferoniego wyznacz odpowiednie oszacowania teoretyczne.

Malgorzata Bogdan