Analiza Dużych Zbiorów Danych

Lista 1 - Wielokrotne testowanie

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

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

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

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.

1. **a)** Dla każdego z powyższych modeli wyznacz estymator najmniejszych kwadratów dla wektora β i wykonaj testy istotności jego elementów.

   b) Porównaj jak się zmienia odchylenie standarowe estymatora β_1 i szerokość 95% przedziału ufności dla tego parametru w miarę tego jak rośnie rozważany model.

   c) Porównaj liczbę prawdziwych i fałszywych odkryć dla różnych modeli.

   d) Porównaj wyniki **1c)** z liczbą prawdziwych i fałszywych 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 i porównaj z wartością teoretyczną (patrz odwrotny rozkład Wisharta).

   b) Średnią szerokość 95% przedziału ufności dla β_1 i porównaj z teoretycznym oszacowaniem.

   c) Średnią liczbę prawdziwych i fałszywych odkryć dla procedur testowania bez korekty oraz z korektą Bonferoniego i BH.

   d) Estymatory FWER i FDR dla procedur testowania bez korekty oraz z korektą Bonferoniego i BH.

   e) Dla procedur bez korekty i z korektą Bonferoniego wyznacz odpowiednie oszacowania teoretyczne średniej liczby fałszywych i prawdziwych odkryć oraz FWER.