Analiza Dużych Zbiorów Danych

Lista 3 - Błąd predykcji i kryteria informacyjne

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

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

gdzie $\beta = (\beta_1, \ldots, \beta_{950})^T$, $\beta_1 = \ldots = \beta_{20} = 3.5$, $\beta_i = 0, i > 20$, $\varepsilon \sim N(0, I)$.

Wykonaj następujące analizy w oparciu o modele wykorzystujące

i) pierwszych 10 zmiennych;

ii) pierwszych 20 zmiennych;

iii) pierwszych 30 zmiennych;

iv) pierwszych 50 zmiennych;

v) pierwszych 100 zmiennych;

vi) pierwszych 500 zmiennych;

vii) wszystkie 950 zmiennych.

1. **a)** Dla każdego z powyższych modeli
   - Wyestymuj β metodą najmniejszych kwadratów i wyznacz $\text{RSS} = ||\hat{Y} - Y||^2$ oraz wylicz oczekiwaną wartość błędu predykcji

     $$\text{PE} = \mathbf{E}_{\varepsilon^*}||Y^* - \hat{Y}||^2, \quad Y^* = X\beta + \varepsilon^*,$$

     gdzie $\varepsilon^* \sim N(0, I)$ jest wektorem losowym niezależnym od próby treningowej.
   - Użyj RSS do estymacji PE wykorzystując prawdziwą wartość i zastępując ją jej klasycznym nieobciążonym estymatorem.
   - Wyestymuj PE stosując walidację krzyżową typu "leave-one-out".

   b) Wybierz optymalny model stosując powyższe estymatory PE.

   c) Powtórz powyższe analizy 100 razy i dla każdego z powyższych modeli porównaj wykresy pudełkowe wartości $\hat{\text{PE}} - \text{PE}$ dla wyżej wymienionych estymatorów PE.

2. Zastosuj AIC, BIC, RIC, mBIC i mBIC2 (można użyć biblioteki *bigstep* w `R`) do identyfikacji istotnych zmiennych w bazach danych składających się z

   i) pierwszych 50 zmiennych;

   ii) pierwszych 100 zmiennych;

   iii) pierwszych 200 zmiennych;

   iv) pierwszych 500 zmiennych;

   v) wszystkich 950 zmiennych.

   a) Podaj liczbę prawdziwych i fałszywych odkryć i kwadratowy błąd estymacji wektora $EY = X\beta$:

   $$\text{SE} = ||X\hat{\beta} - X\beta||^2.$$

   b) Powtórz punkt a) 100 razy i podaj wyestymowaną moc, FDR i średni błąd kwadratowy estymacji $EY = X\beta$ dla wszystkich powyższych kryteriów. Omów uzyskane wyniki względem liczby predyktorów.

3. Powtórz zadania 1 i 2 w sytuacji gdy $n = 5000$. Omów uzyskane wyniki względem liczby obserwacji.