SAR 2020/2021 Laboratorium 7

17-18.11.2020

9.1 (selekcja zmiennych BIC, AIC)

Dane w pliku *uscrime.txt* zawierają informacje dotyczące 47 stanów USA:

R - wskaźnik przestępczości, S - =1 (stany południowe), = 0 (pozostałe stany),

Age - liczba męższczyzn w wieku 14-24 przypadających na 1000 mieszkańców,

Ex0, Ex1 - wydatki na policję w latach, odpowiednio, 1960 i 1959,

LF - wskaźnik udziału pracowników w wieku 14-24 lat, W - wskaźnik dobrobytu,

M - liczba mężczyzn przypadających na 1000 kobiet, N - liczba mieszkańców stanu (w setkach tys.),

NW - liczba osób rasy niebiałej przypadających na 1000 mieszkańców,

U1, U2 - wskaźnik bezrobocia dla mężczyzn w wieku, odpowiednio, 14-24 lat i 35-39 lat,

X - wskaźnik nierówności dochodu (liczba rodzin na 1000, których dochód jest mniejszy niż połowa mediany dochodu wszystkich rodzin).

a) Dopasować model opisujący zależność współczynnika przestępczości od pozostałych zmiennych. Wykonaj wykresy rozproszenia i obliczyć współczynniki korelacji dla wszystkich par zmiennych . Znaleźć parę zmiennych najsilniej skorelowanych i usunąć jedną z tych zmiennych z modelu.

b) Wybrać „najlepszy" podzbiór zmiennych objaśniających stosując:
· metodę pełnego przeszukiwania przestrzeni modeli (napisz własną funkcję w R),
· kryteria: AIC, BIC, modyfikowany R^2.

c) Wybrać „najlepszy" podzbiór zmiennych objaśniających stosując:
· metody: eliminacji (selekcja wstecz), dołączania (selekcja wprzód), selekcję krokową,
· kryteria: AIC, BIC, modyfikowany R^2.

d) Wybrać „najlepszy" podzbiór zmiennych objaśniających stosując metodę opartą na wstępnym uporządkowaniu zmiennych według t-statystyk. Metoda działa w następujący sposób. Dopasowujemy model pełny i obliczamy t-statystyki dla wszystkich zmiennych. Następnie porządkujemy zmienne według t-statystyk (od najbardziej istotnej do najmniej istotnej). Z zagnieżdżonej rodziny modeli (danej przez uporządkowanie) wybieramy ten dla którego wartość kryterium BIC/AIC jest minimalna.

9.2 (BIC, AIC)

Wykonaj następujący eksperyment. Celem eksperymentu jest porównanie jak zmienia się prawdopodobieństwo poprawnej selekcji w zależności od wielkości próby dla kryteriów BIC i AIC.

a) Wygeneruj dane (X, Y) zakładając że wiersze macierzy X są generowane z $p = 9$- wymiarowego rozkładu normalnego (zakładamy że zmienne są niezależne) według równania liniowego:

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

gdzie $\beta = (1, 1, 1, 0, 0, 0, 0, 0, 0)'$ i ϵ jest wektorem błędów z rozkładu standardowego normalnego. Uwaga: nie ma wyrazu wolnego!

b) Użyj funkcji `step()` z argumentem `direction=''backward''` aby wybrać prawdziwy model, t.j. model składający się z pierwszych trzech zmiennych.

c) Powtórz powyższą procedurę $L = 50$ razy aby wyestymować prawdopodobieństwo poprawnej selekcji

d) Powtórz eksperyment dla $n = 25, 50, 75, 100, 125, 150, 175, 200$ dla AIC i BIC. Zaprezentuj otrzymane wyniki na wykresie pokazującym zależność prawdopodobieństwa poprawnej selekcji od wielkości próby n.

Podpowiedź: przydatna funkcja w R: {`setequal()`}. `\ Uwaga: obliczenia mogą zająć parę minut.`