

**6.1** Wczytaj dane *longley*. Zbiór zawiera informacje o liczbie osób zatrudnionych w USA w latach 1947-1962.

(a) Dopasuj model liniowy opisujący zależność liniową między zmienną Employed a pozostałymi zmiennymi.

(b) Oblicz:

- macierz korelacji pomiędzy zmiennymi,
- współczynniki determinacji wielokrotnej dla zmiennych objaśniających,
- współczynniki podbicia wariancji dla zmiennych objaśniających,

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

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

$Age$  - liczba mężczyzn 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) Dopasuj model opisujący zależność współczynnika przestępczości od pozostałych zmiennych. Sporządź wykresy rozproszenia i oblicz współczynniki korelacji dla wszystkich par zmiennych objaśniających. Znajdź parę zmiennych najsilniej skorelowanych i usuń jedną z tych zmiennych z modelu.

(b) Wybierz "najlepszy" podzbiór zmiennych objaśniających stosując:

- metodę pełnego przeszukiwania przestrzeni modeli (napisz własną funkcję w R),
- kryteria: AIC, BIC, modyfikowany  $R^2$ .

(c) Wybierz "najlepszy" podzbiór zmiennych objaśniających stosując:

- metody: eliminacji (selekcja wstecz), dołączania (selekcja wprzód), selekcję krokową,
- kryteria: AIC, BIC, modyfikowany  $R^2$ .

(d) Wybierz "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.

### 6.3

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

- 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  $\epsilon$  jest wektorem błędów z rozkładu standardowego normalnego. Uwaga: nie ma wyrazu wolnego!

- Użyj funkcji `step()` z argumentem `direction="backward"` aby wybrać prawdziwy model, t.j. model składający się z pierwszych trzech zmiennych.

- Powtórz powyższą procedurę  $L = 50$  razy aby wyestymować prawdopodobieństwo poprawnej selekcji
- 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.