SAR 2020/2021 Laboratorium 9

1-2.12.2020

9.1 RIDGE

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

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

b) Obliczyć:

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

c) Obliczyć estymatory parametrów używając metody regresji grzbietowej korzystając z definicji oraz rozkładu SVD.

d) Dopasować model regresji grzbietowej korzystając z pakietu `glmnet`. Wybrać optymalną wartość parametru λ stosując metodę kroswalidacji. Zbadać stabilność estymacji dla różnych wartości parametru λ. Obliczyć estymatory współczynników dla $\lambda = 0.03$.

e) Porównać współczynnik przy zmiennej GNP (Produkt narodowy brutto) dla modelu z punktu (a) oraz dla modelu dopasowanego przy pomocy regresji grzbietowej z parametrem $\lambda = 0.03$.

9.2 LASSO, selekcja

Wczytać dane z pliku *prostate.data*. Dane zawierają informacje dotyczące raka prostaty u 97 mężczyzn.

a) Dopasować model liniowy opisujący zależność zmiennej lpsa (logarithm of prostate specific antigen) od pozostałych zmiennych (z wyjątkiem zmiennej train). Dokonać selekcji zmiennych metodą eliminacji "wstecz", z kryterium AIC.

b) Dopasować model używając metody lasso. Przeanalizować zachowanie estymatorów w zależności od parametru λ.

c) Na podstawie metody lasso i kroswalidacji, dokonać selekcji zmiennych. Wyświetlić które zmienne są po kolei dołączane do modelu gdy rośnie wartość parametru λ.

d) Obliczyć wartości estymatorów w wybranym modelu.

e) Losowo wybrać 80 obserwacji i dopasować do nich model MNK, MNK z lasso wybranym kroswalidacyjnie i MNK do zmiennych wybranych przez lasso. Porównać współczynniki w modelach i MSE (Błąd średniokwadratowy) predykcji dla pozostałych 17 obserwacji.

9.3 RIDGE wariancja

Udowodnij że:

a) Estymator $\hat{\beta}^{RIDGE}$ jest obciążony.

b) Estymator RIDGE ma mniejszą wariancję niż estymator MNK, tzn.:

$$var(\hat{\beta}^{RIDGE}) \leq var(\hat{\beta}^{MNK}).$$

Zaproponuj eksperyment symulacyjny, który ilustruje powyższy wynik teoretyczny.