Lista 1

Aneta Przydróżna

Niech $X_{1000\mathbf{x}950}$ będzie macierzą niezależnych zmiennych losowych z rozkładu $N\left(0, \frac{1}{\sqrt{1000}}\right)$, a $\beta = (3,3,3,3,3,0,...,0)^T$ i wektor szumu $\epsilon \sim N(0,I)$.

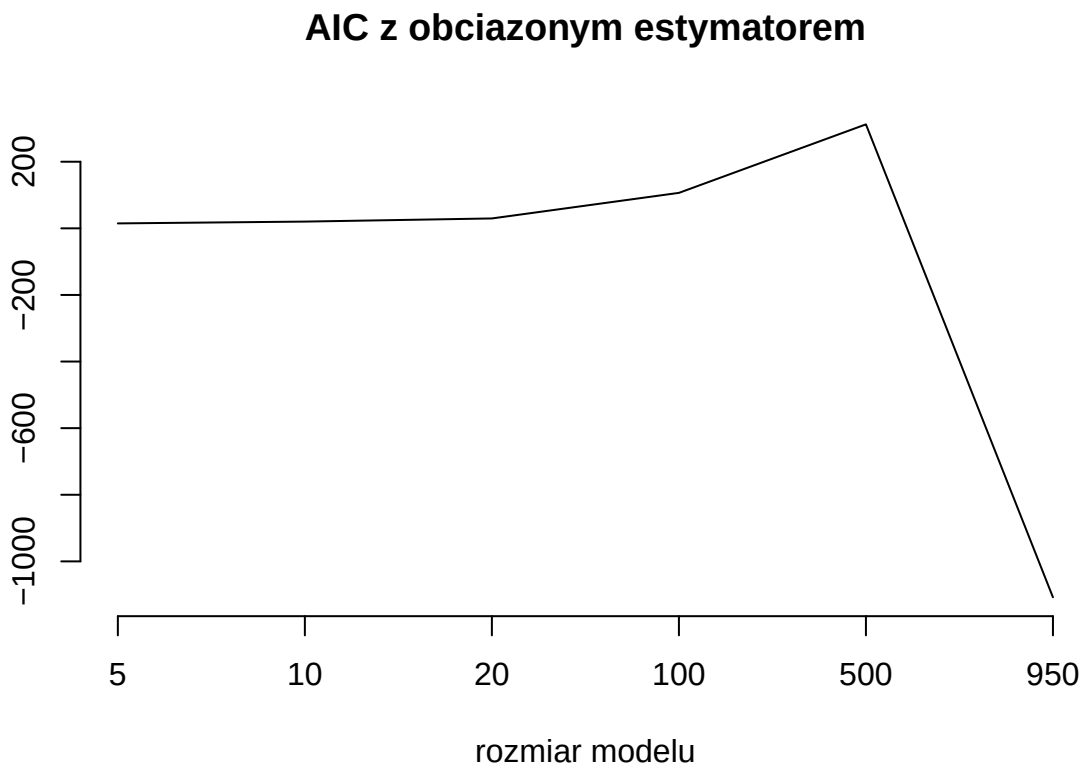
Wybór optymalnego modelu macierzy planu $X_s \subseteq X$, gdzie S={5,10,20,100,500,950}.

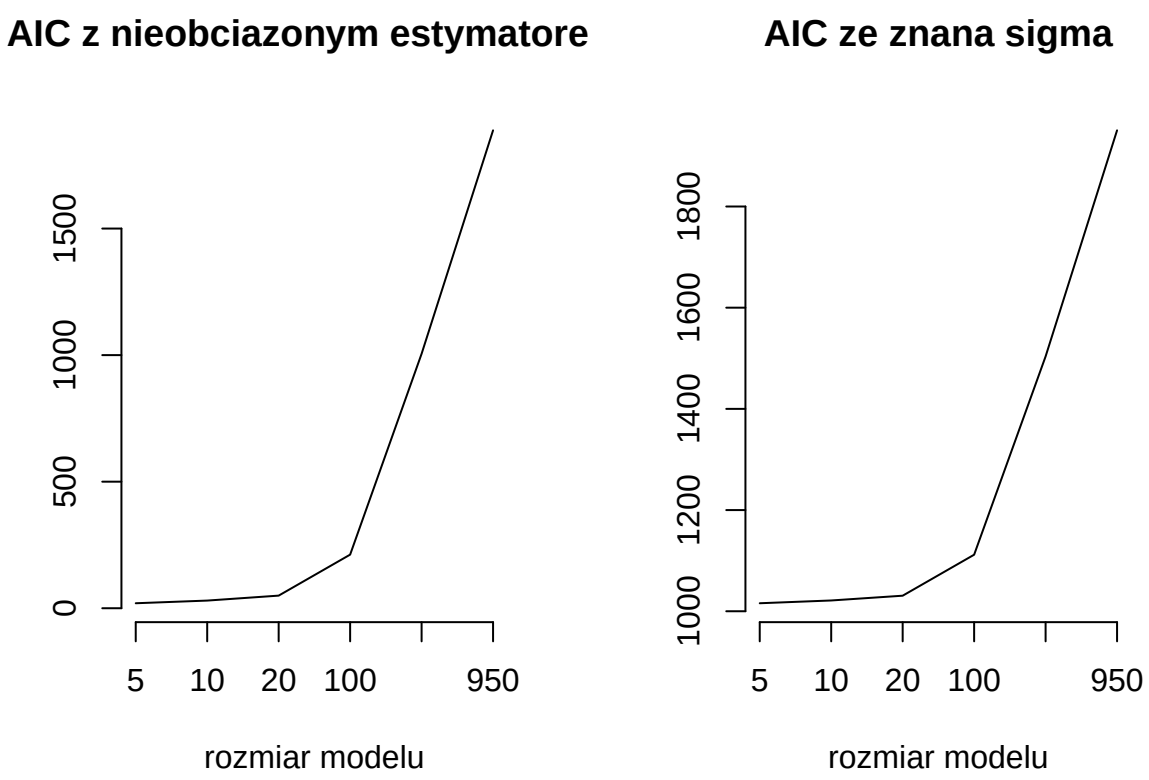
• Estymacja β metodą najmniejszych kwadratów $\hat{\beta(s)} = (X_s^{'}X_s)^{-1}X_s^{'}Y$.

• Oczekiwany błąd predykcji $PE = E||X(\beta - \hat{\beta}) + \epsilon^*||^2$, gdzie $\epsilon^* \sim N(0,I)$ jest nowym szumem niezależnym od ϵ.

• Estymacja PE przy założeniu, że σ jest znana oraz zastępując ją nieobciążonym $\sigma^2 = \frac{RSS}{n-p}$ i obciążonym $\sigma^2 = \frac{RSS}{n}$ estymatorem.

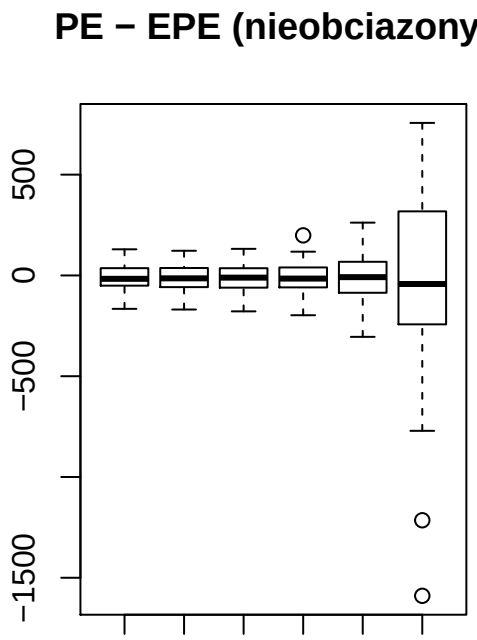
• Estymacja PE przy pomocy walidacji krzyżowej $\hat{PE} = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)$, gdzie $e_i = Y_i - X_i\hat{\beta_{ols}}$ i $h_{ii} = X_i(X_i^{'}X_i)^{-1}X_i^{'}$.

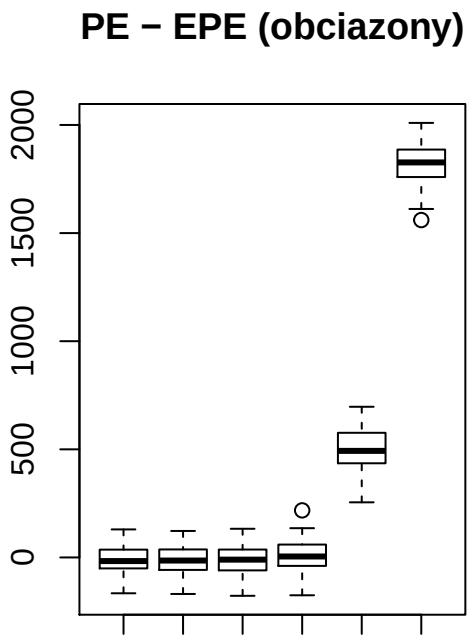


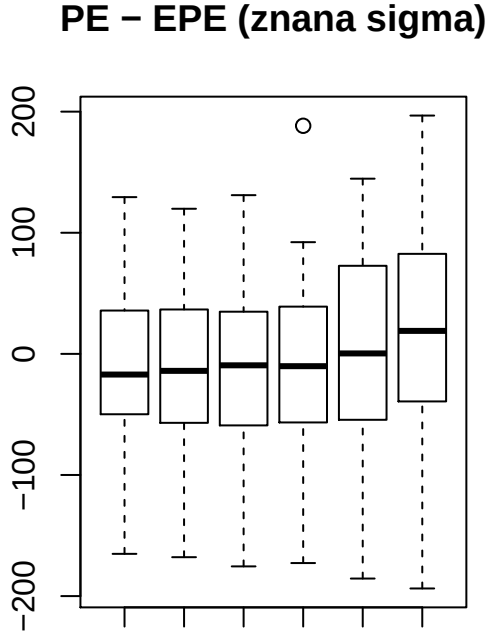


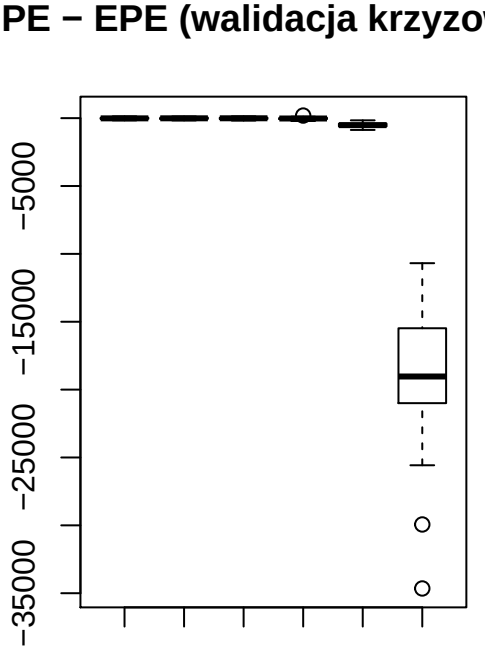
Wykresy kryterium AIC do s=20, powoli rosną, do s=100 trochę szybciej, a do s=950 AIC dla modelu ze znaną σ i nieobciążonym estymatorem, bardzo szybko rosną do 1000, natomiast dla modelu z obciążonym estymatorem, po przekroczeniu s=500, czyli w przypadku gdy liczba parametrów przekracza połowę liczby obserwacji, wykres gwałtownie zaczyna spadać poniżej 0. Jednak we wszystkich przypadkach najlepsze wyniki kryterium odpowiadają s=5.

Poniżej znajdują się boxploty $\hat{PE} - PE$, dla czterech metod estymacji PE, powtórzonego 100-krotnie doświadczenia.









Boxploty różnic dla nieobciążonego estymatora σ stopniowo zwiększają rozrzut w granicach (-400,400),lecz dla ostatniego modelu z liczbą parametrów, zbliżającą się do liczby obserwacji, przeciał ten jest trzykrotnie szerszy. Dla obciążonego estymatora σ, cztery pierwsze boxploty są podobne, gdy s=500 widać, że oczekiwana wartość błędu predykcji jest zawsze większa od estymowanej, wartości odstających jest mało, a różnice sięgają 1000, dla s=950 podobnie ale dwukrotnie więcej. W przypadku znanej σ wszystkie boxploty są podobne, z wieloma wartościami odstającymi rozrzuconymi w granicach (-200,200). Można wywnioskować, że wielkość macierzy planu nie ma dużego wpływu na estymacje błędu predykcji. Policzony metodą walidacji krzyżowej błąd nie różni się znacząco od prawdziwego w czterech pierwszych modelach, dla ostatniego różnica jest ujemna, sięgająca -40000, zatem ta medoda nie radzi sobie z dużą liczbą parametrów.