ADZD - lista 2

Porównanie metod estymacji średniej wielowymiarowego rozkładu normalnego pod kątem minimalizacji średniego błędu kwadratowego.

Wstęp

Załóżmy, że dysponujemy zbiorem obserwacji $X_1, ...X_p$ - niezależnych zmiennych losowych z rozkładu normalnego o znanej wspólnej wariancji σ^2 i nieznanych, potencjalnie różnych średnich $\mu_1, ...\mu_p$, które chcemy estymować. Estymator największej wiarygodności jest wektorem średnich próbkowych $\hat{\mu}_{LS} = (\bar{X}_1, ...\bar{X}_p) = (X_1, ...X_p)$. Spośród estymatorów nieobciążonych jest on najlepszy pod kątem minimalizacji średniego błędu kwadratowego.

$$MSE(\hat{\mu}_{LS}) = E||(\hat{\mu}_{LS} - \mu)||_2^2 = ... = p\sigma^2$$

Możemy zredukować MSE zastępując $\hat{\mu}_{LS}$ jednym z opisanych niżej estymatorów obciążonych.

Estymatory Jamesa-Steina

Spróbujemy skonstruować estymator $\hat{\mu}_c = c\hat{\mu}_{LS}$, gdzie c jest stałą dobraną tak by minimalizować MSE. Wtedy

$$c = \frac{||\mu||^2}{||\mu||^2 + p\sigma^2} \in [0, 1)$$

$$MSE(\hat{\beta}_c) = c \cdot p\sigma^2 \leq p\sigma^2$$

W ten sposób uzyskujemy estymator lepszy niż $\hat{\mu}_{LS}$, jednak niemożliwy do zastosowania w praktyce ze względu na to, że nie znamy prawdziwej wartości μ. Odpowiedzią na ten problem jest korekta wprowadzona przez Jamesa i Steina (1961r.) polegająca na zastąpieniu nieznanego c przybliżającym je wyrażeniem:

$$c_{JS} = 1 - \frac{(p-2)\sigma^2}{||\hat{\mu}_{LS}||^2}$$

Estymator ten nazywany jest **ściągającym do zera**. Korzystając z twierdzenia Steina można pokazać, że teoretyczna wartość MSE dla estymatora Jamesa-Steina ściągającego do zera to

$$E||\hat{\mu}_{c_{JS}} - \mu||^2 = p\sigma^2 - \sigma^4 \frac{(p-2)^2}{||\mu_{LS}||^2} < p\sigma^2, \text{ gdy } p > 2.$$

Analogicznie konstruujemy **estymator ściągający do wspólnej średniej** szukając stałej d takiej, by minimalizować MSE estymatora $\hat{\beta}_d = (1-d)\hat{\mu}_{LS} + d\bar{X}$ i uzyskujemy

$$d = \frac{\sigma^2}{Var(\mu) + \sigma^2},$$

$$MSE(\hat{\beta}_d) = \frac{p\sigma^2 Var(\mu) + \sigma^4}{Var(\mu) + \sigma^2} \leq p\sigma^2,$$

gdzie $Var(\mu) = \frac{1}{p-1}||\mu - \bar{\mu}||^2$. Wyrażenie d z nieznaną nam wariancją możemy zastąpić przez

$$d_{JS} = \frac{(p-3)\sigma^2}{(p-1)Var(\mu_{LS})}$$

otrzymując w ten sposób drugi z estymatorów Jamesa-Steina.

Estymator powstały przez twarde ucięcie

Wybieramy procedurę testowania istnotości (np. Bonferroniego) i defniujemy następujący estymator:

$$\hat{\mu} = \begin{cases} X_i, \text{ gdy procedura odrzuciła } H_0 : \beta_i = 0 \\ 0, \text{ gdy nie mamy podstaw do odrzucenia.} H_0 : \beta_i = 0. \end{cases}$$

Estymator ten nazywamy "twardym ucięciem" estymatora $\hat{\mu}_{LS}$. W przypadku procedur Bonferroniego i Benjamini'ego-Hochbegra $MSE(\hat{\mu}) \to_p MSE_{optymalne}$.

Estymator Bayesowski

Tym razem rozważamy problem testowania hipotez postaci

$$H_0 : X_i \sim P_0 \quad \text{vs} \quad H_1 : X_i \sim P_1,$$

gdzie P_0 i P_1 są rozkładami o gęstościach odpowiednio f_0, f_1. *Funkcja straty* c to funkcja która każdej parze ("stan faktyczny - s", "decyzja testu - d") przyporządkowuje wartość zgodnie z następującymi zasadami:

- jeśli decyzja jest słuszna, $c(s, d) = 0$,
- jeśli robimy błąd typu I $c(s, d) = c_0$,
- jeśli robimy błąd typu II $c(s, d) = c_1$.

Testowanie każdej hipotezy będzie polegało na odrzuceniu H_0, i, gdy statystyka testowa $X_i \in R$, gdzie R nazywamy *obszarem odrzuceń*. **Rozważany przez nas estymator to taki, dla którego wybór obszaru R minimalizuje wartość oczekiwaną straty.** Dysponując dodatkowo informacją o tym jak prawdopodobne jest że $H_{0,i}$ jest prawdziwa ($P(H_0)$) możemy wyznaczyć R w następujący sposób:

$$E[c] = E[c|H_0]P(H_0) + E[c|H_1](1 - P(H_0))$$
$$E[c|H_0] = 0 \cdot (1 - P(I)) + c_0 P(I)$$

$$E[c|H_1] = 0 \cdot (1 - P(II)) + c_1 P(II)$$

Za $P(I), P(II)$ postawimy odpowiednio całki $\int_R f_0(x)dx$ i $1 - \int_R f_1(x)dx$ otrzymując w ten sposób

$$E[c] = P(H_0) \cdot c_0 \int_R f_0(x)dx + P(H_1) \cdot c_1 \left(1 - \int_R f_1(x)dx\right) = c_1 + \int_R c_0 f_0(x)P(H_0) - c_1 f_1(x)P(H_1)dx.$$

Dla ustalonych $c_0, c_1, P(H_0), P(H_1), f_0(x), f_1(x)$ wyrażenie przyjmuje najmniejszą wartość, gdy całka jest jak najbardziej ujemna, a więc gdy wybieramy maksymalny obszar R dla którego wyrażenie podcałkowe jest ujemne. Otrzymamy w ten sposób

$$R = \{x : \frac{f_0(x)}{f_1(x)} < \frac{c_1 P(H_1)}{c_0 P(H_0)}\}$$

Zadanie 1

Estymator współczynników regresji liniowej otrzymany metodą najmniejszych kwadratów dany jest wzorem $\hat{\beta}_{LS} = (X^T X)^{-1} X^T Y$ i ma rozkład normalny $N(\beta, \sigma^2 (X^T X)^{-1})$, mamy więc do czynienia z problemem estymacji średnich w wielowymiarowym rozkładzie normalnym. W przypadku, gdy macierz X jest ortogonalna, wyrażenie $(X^T X)^{-1} = I$ znika a estymator jest postaci

$$\hat{\beta}_{OLS} = X^T Y$$

Zakładając, że wariancja błędów losowych jest znana możemy obliczyć również p-wartości dla testów istotności $\hat{\beta}_{LS}$, które przydadzą się w zadaniu 3. Wiemy, że przy hipotezie zerowej $\hat{\beta}_{LS} = \hat{\beta}$ ma rozkład normalny ze średnią 0 i wariancją σ^2, zatem p-wartość wyliczymy ze wzoru

$$pval(\hat{\beta}_i) = 2(1 - \Phi_{N(0,\sigma^2)}(|\hat{\beta}_i|)).$$

Zadanie 2

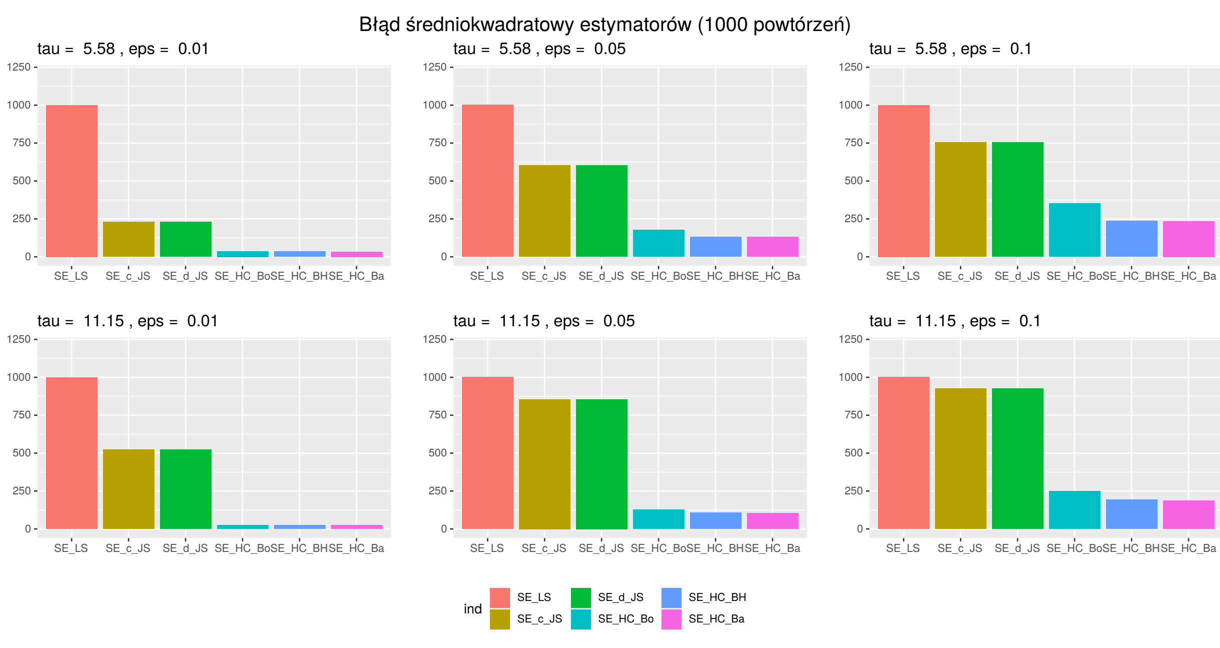
Estymatory zostały wyznaczone we wstępie.

Zadanie 3

W każdej replikacji eksperymentu generujemy wektor współczynników "prawdziwego" modelu β tak, że element β_i z prawdopodobieństwem ϵ pochodzi z rozkładu normalnego z wariancją τ^2, a z $1 - \epsilon$ jest zerem. Porównamy wyniki dla różnych wartości tych 2 parametrów.

Wykresy: Błąd średniokwadratowy

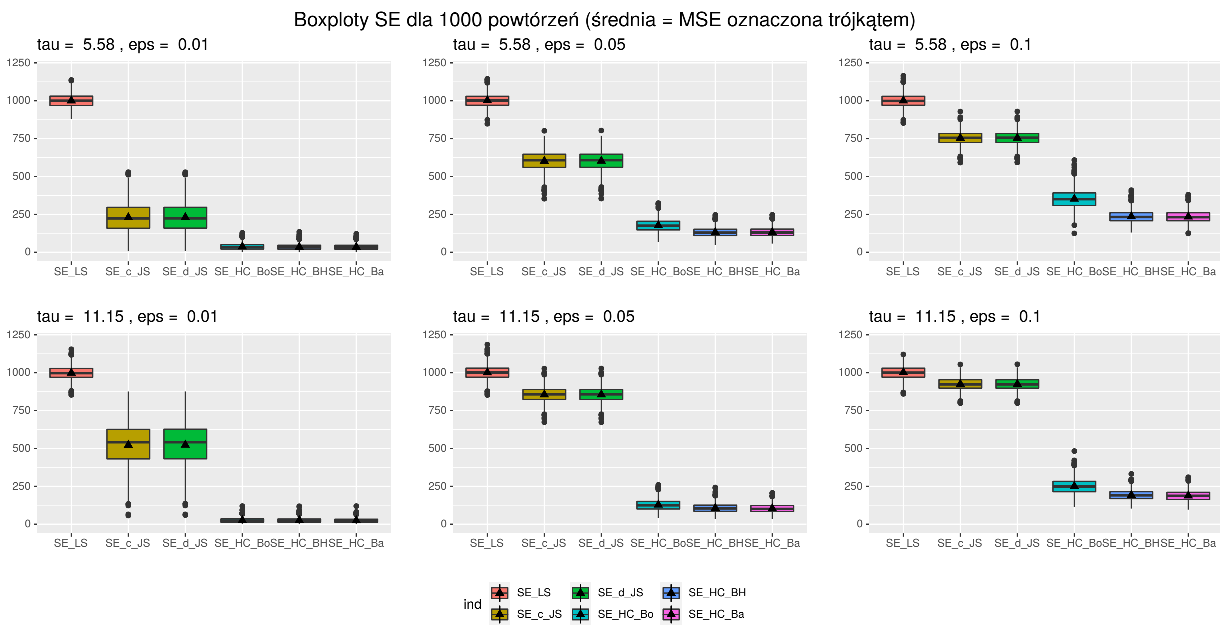
Wykres poniżej przedstawia kwadratowy błąd predykcji uśredniony dla 1000 replikacji eksperymentu.



Komentarz:

- MSE estymatora $\hat{\beta}_{LS}$ niezależnie od τ, ϵ wynosi około 1000, zgodnie z teoretycznymi założeniami ($MSE(\hat{\beta}_{LS}) = p\sigma^2 = 1000$). Równocześnie jest wyższe niż MSE innych estymatorów.
- estymatory Jamesa-Steina mają MSE na bardzo podobnym poziomie, znacznie lepsze niż estymator najmniejszych kwadratów.
- estymatory powstałe przez twarde odcięcie są jeszcze lepsze - dla wszystkich procedur wyniki są zbliżone. Można podejrzewać, że estymatory powstałe przez twarde odcięcie są najlepsze, bo w naszej symulacji mamy do czynienia z danymi z podobnego modelu, w którym β_i są zupełnie zerowe.
- MSE wszystkich estymatorów oprócz LS rośnie wraz ze wzrostem τ - czyli i ze wzrostem ϵ. Większy ϵ oznacza, że prawdziwy wektor β ma mniej zer, z kolei większe τ to większa wariancja elementu β_i. Znaczy to, że łatwiej jest estymować współczynniki w modelu z małą liczbą niezerowych β_i o małej wariancji.

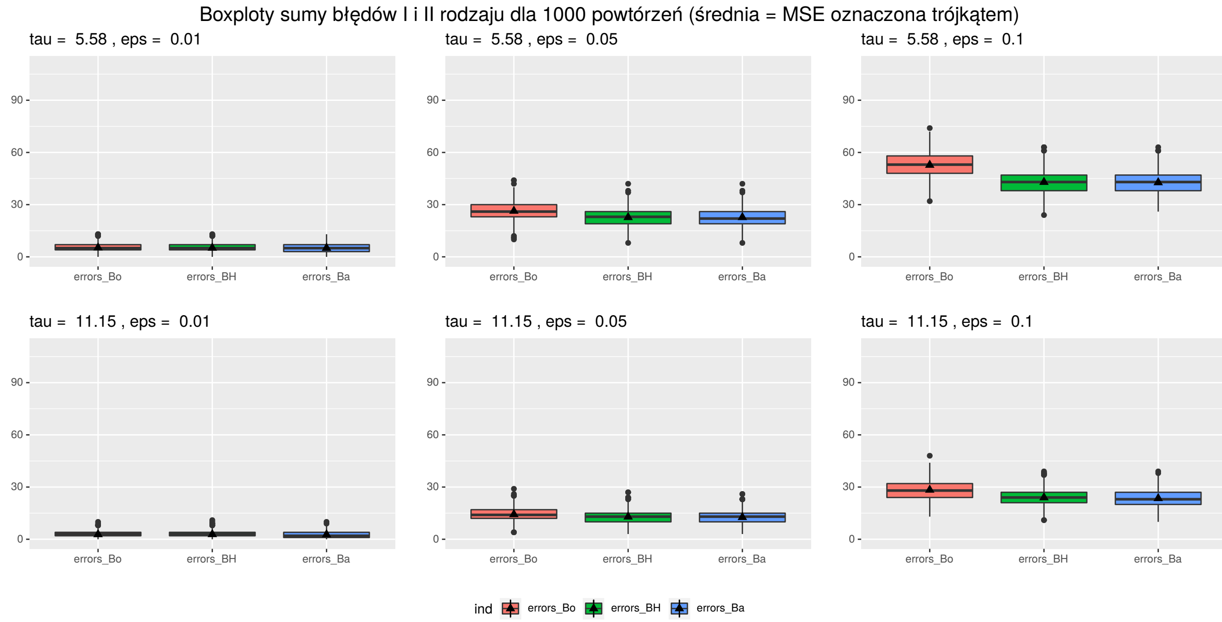
Poniżej dodatkowe wykresy pudełkowe błędów kwadratowych rozważanych estymatorów:



Komentarz: Można zauważyć pewien wpływ parametrów τ i ϵ na wariancję estymatorów - najbardziej widoczny jest on w przypadku estymatorów Jamesa-Steina. W ich przypadku rozrzut PE na wykresach maleje wraz ze wzrostem ϵ (odsteka niezerowych β_i) - dla pozostałych estymatorów jest odwrotnie. Wydaje się, że parametr τ nie wpływa na wariancję pozostałych estymatorów, a wpływa na wariancję estymatorów Jamesa-Steina (większe τ to większa wariancja $MSE(\mu_{JS})$.

Wykresy: sumy błędów dla procedur testowania

Na wykresach porówniujemy sumy błędów I i II rodzaju dla procedury Bonferroniego, Benjaminiego-Hochbergra i estymatora Bayesa.



Komentarz:

- Suma błędów I i II rodzaju jest na podobnym poziomie zarówno pod względem średniej, jak i wariancji.
- Można zauważyć wpływ parametrów eksperymentu na sumy błędów - im większy jest ϵ, tym więcej jest błędów dla każdej procedury. W przypadku τ zależność jest odwrotna i mniej widoczna.
- Procedura Bonferroniego jest trochę gorsza niż 2 pozostałe.

Pominęłam podpunkty a. i b. w zadaniu, gdyż ich wyniki (tzn. wyniki dla tylko 1 replikacji) w jakiś sposób zawierają się w wykresach pudełkowych.