Raport 2 — Analiza dużych zbiorów danych

Magdalena Potok

2024-04-24

Celem raportu jest porównanie estymatorów β pod kątem błedu kwadratowego oraz porcedur testowania porównując liczbę błedów I i II rodzaju.

Wygenerowana została ortonormalna macierz planu $\mathbb{X}_{1000\times1000}$, a następnie wektor współczynników regresji jako ciąg niezależnych zmiennych losowych z rozkładu

$$\beta_i \sim (1-\gamma)\delta_0 + \gamma\phi(0,\tau^2),$$

gdzie δ_0 jest rozkładem skupionym w zerze, $\phi(0,\tau^2)$ jest gęstością rozkładu $N(0,\tau^2)$ i mamy 6 przypadków, gdy $\gamma \in \{0.01, 0.05, 0.1\},\ \tau \in \{1.5\sqrt{2log1000}, 3\sqrt{2log1000}\}$. Dla kazdego z przypadków wygenerowany został wektor odpowiedzi $Y = \mathbb{X}\beta + \epsilon$, gdzie $\epsilon \in N(0, I_{1000\times1000})$. Zakładamy, że wariancja błędu jest znana $\sigma^2 = 1$.

Zadanie 1

Estymator najmniejszych kwadratów $\hat{\beta}_{LS}$ dla wektora odpowiedzi β jest również estymatorem największej wiarygodności ozn. $\hat{\beta}_{MLE}$. Na ogół jest on zadany wzorem $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$ i pochodzi z rozkładu $N(\beta, \sigma^2(X'X)^{-1})$. Przy założeniach macierzy planu i wariancji błedu wypisanych wyżej możemy uprościć tę postać i dostajemy, że

$$\hat{\beta}_{LS} = X'Y,$$

ten estymator pochodzi z rozkładu $N(\beta, I)$. Kolejne poznane estymatory to estymatory Jamesa-Steina, korzystają one jednak z estymatora największej wiarogodności. **Estymator Jamesa-Steina ściągający do zera** dla parametru β jest postaci:

$$\hat{\beta}_{JS_1} = c_{JS}\hat{\beta}_{MLE}, \quad c_{JS} = 1 - \frac{(n-2)\sigma^2}{||\hat{\beta}_{MLE}||^2}.$$

Kolejnym estymatorem jest **Estymator Jamesa-Steina ściągajacy do wspólnej średniej** i jest on postaci:

$$\hat{\beta}_{JS_2} = (1-d_{JS})\hat{\beta}_{MLE} + d_{JS}\bar{\hat{\beta}}_{MLE}, \quad d_{JS} = \frac{n-3}{n-1}\frac{1}{Var(\hat{\beta}_{MLE})}.$$

Zadanie 2

Poniżej przedstawione zostaną indeksy istotnych zmiennych, gdy $\gamma = 0.01,\ \tau = 1.5\sqrt{2log1000}$ dla różnych procedur.

a) procedura Bonferroniego: 96, 376

b) procedura Benjaminiego-Hochberga: 376, 96, 655, 249, 12

c) klasyfikator Bayesowski: 12, 96, 249, 376, 655.

Prawdziwymi istotnymi współczynnikami w tym przypadku są: 10, 12, 96, 376, 655. Możemy zauważyć, że porcedura Bonferroniego jest najbardziej konserwatywna, ma najmniejszą liczbę odkryć.
Pozostałe przypadki dla innych γ oraz τ zostaną przeanalizowane pod kątem sumy błędów I i II rodzaju w zadaniu 4.

Zadanie 3

Dla każdej procedury z zadania 2. wyznaczone zostały „ucięte” estymatory wektora β za pomocą poniższego kodu.

```
uciete_estym <- function(indeks){ #indeks istotnych wspolczynnikow
  indx <- rep(0,1000)
  indx[indeks] < -1
  return(indx)
}
indx_BF <- mle_beta * uciete_estym(discovery_BF)
indx_BH <- mle_beta * uciete_estym(discovery_BH)
indx_Bay <- mle_beta * uciete_estym(discovery_bayes)
```

Zadanie 4

Estymatory z zadania 1. i 3. porównane zostały pod kątem błędu kwadratowego.

Tabela 1: Błąd kwadratowy estymatorów.

γ	τ	$\hat{\beta}_{LS}$	$\hat{\beta}_{c_{JS}}$	$\hat{\beta}_{d_{JS}}$	$\hat{\beta}^{uc}_{BF}$	$\hat{\beta}^{uc}_{BH}$	$\hat{\beta}^{uc}_{Bay}$
0.01	$1.5\sqrt{2\log 1000}$	955.959	367.132	367.377	33.072	33.072	33.072
0.05	$1.5\sqrt{2\log 1000}$	1047.282	476.432	475.729	33.039	48.018	48.018
0.10	$1.5\sqrt{2\log 1000}$	1012.131	579.188	579.394	176.481	113.567	113.567
0.01	$3\sqrt{2\log 1000}$	977.579	843.720	845.133	110.514	77.318	91.779
0.05	$3\sqrt{2\log 1000}$	1002.413	755.667	755.338	423.214	251.249	251.249
0.10	$3\sqrt{2\log 1000}$	975.658	893.223	893.819	183.974	171.338	176.888

Z tabeli można odczytać, że estymatory wyznaczone w zadaniu 3. spisują się najlepiej — mają dużo mniejsze wartości błędu kwadratowego. Najmniejsze wartości dla jednego powtórzenia w każdym z przypadków uzyskał „ucięty” estymator wyznaczony za pomocą procedury Benjaminiego-Hochberga.
Dla każdej procedury testowania z 2. zadania przedstawione zostaną sumy liczb błędów I i II rodzaju.

Tabela 2: Suma liczb błędów I i II rodzaju.

γ	τ	BF	BH	Bayess
0.01	$1.5\sqrt{2\log 1000}$	7	7	7
0.05	$1.5\sqrt{2\log 1000}$	3	4	4
0.10	$1.5\sqrt{2\log 1000}$	32	26	26
0.01	$3\sqrt{2\log 1000}$	13	9	10
0.05	$3\sqrt{2\log 1000}$	55	39	39
0.10	$3\sqrt{2\log 1000}$	33	29	31

Wartości są bardzo do siebie zbliżone, w każdym z przypadków, dla każdej procedury. Cieżko wybrać która z nich wypada najlepiej dla jednego powtórzenia, można zauważyć, że procedura Bonferroniego ma najwięcej błedów i wypada najgorzej.

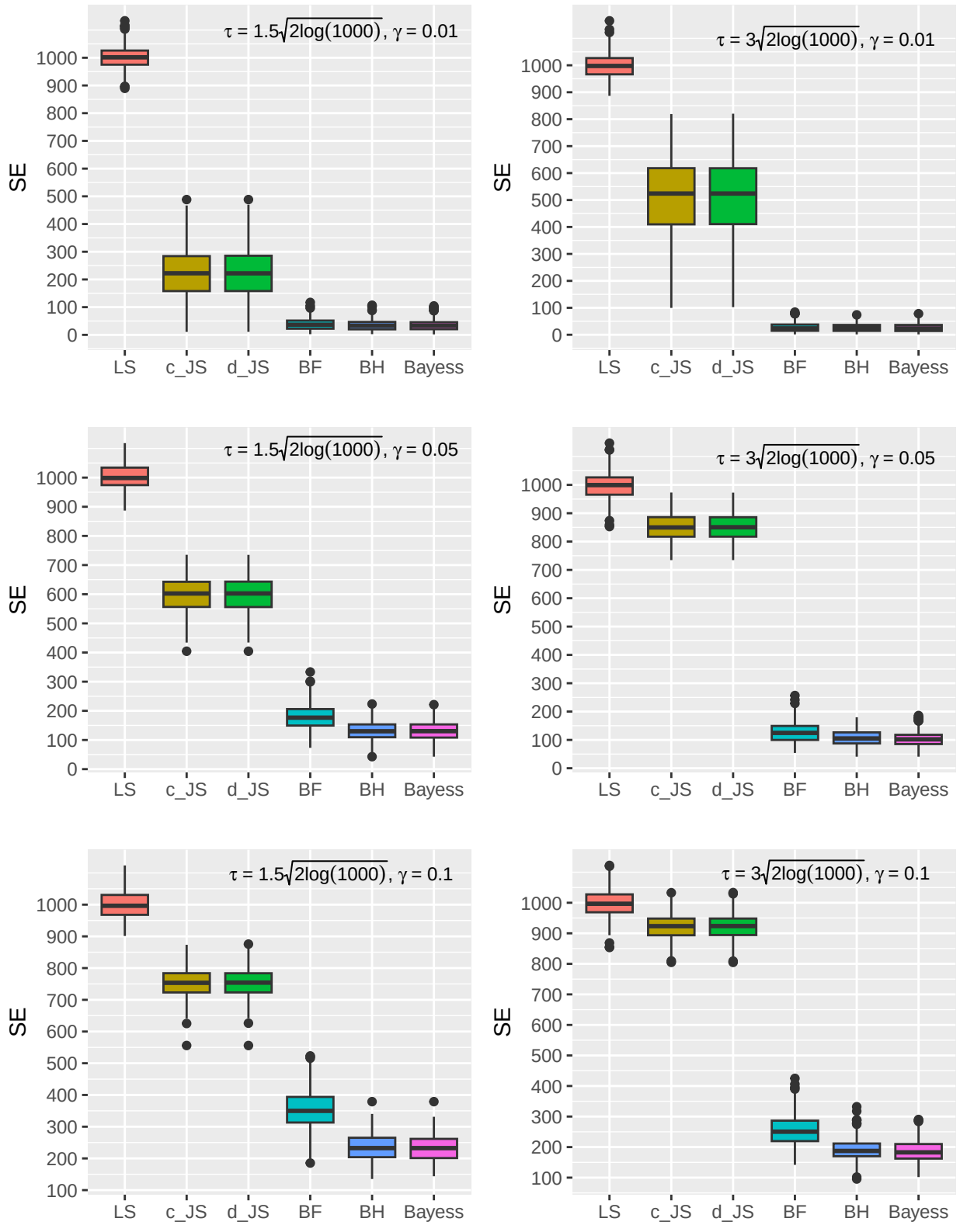
Dla każdej kombinacji γ i τ powtórzę doświadczenie 1000 razy i porównam estymatory pod kątem MSE, a analizowane procedury pod katęm średniej liczby sumy błędów pierwszego i drugiego rodzaju.

Tabela 3: Błąd średniokwadratowy estymatorów przy 1000 powtórzeniach.

γ	τ	$\hat{\beta}_{LS}$	$\hat{\beta}_{c_{JS}}$	$\hat{\beta}_{d_{JS}}$	$\hat{\beta}^{uc}_{BF}$	$\hat{\beta}^{uc}_{BH}$	$\hat{\beta}^{uc}_{Bay}$
0.01	$1.5\sqrt{2\log 1000}$	1000.83	224.81	225.50	34.73	33.21	32.94
0.05	$1.5\sqrt{2\log 1000}$	1001.56	521.53	521.94	25.25	26.46	25.06
0.10	$1.5\sqrt{2\log 1000}$	996.89	596.56	596.83	176.66	130.04	130.66
0.01	$3\sqrt{2\log 1000}$	999.73	853.08	853.25	126.52	106.45	102.52
0.05	$3\sqrt{2\log 1000}$	1005.17	760.31	760.55	355.77	234.71	233.87
0.10	$3\sqrt{2\log 1000}$	1000.45	923.74	923.84	248.30	191.27	186.99

W tabeli zostały wyznaczone średnie wartości błędów kwadratowych estymatorów przy 1000 powtórzeniach. Tak jak przy 1 powtórzeniu „ucięte" estymatory ponownie wypadają dużo lepiej od estymatora najmniejszych kwadratów oraz estymatorów Jamesa-Steina. Ich wartości są bardzo zbliżone dla każdej kombinacji τ i γ. Dla $\tau = 3\sqrt{2log1000}$ „ucięty" estymator wyznaczony za pomocą procedury Bonferroniego minimalnie wypada gorzej, ale nadal te wartości są zbliżone. Estymatory z zadania 2. wypadają znacząco gorzej, szczególnie estymator najmniejszych kwadratów.

Na następnej stronie (rysunek 1.) przedstawione zostały wykresy boxplot dla błędu kwadratowego dla każdej procedury przy 1000 powtórzeniach. Możemy zauważyć z nich, że estymator najmniejszych kwadratów (oznaczony kolorem czerwonym) nie zmienia się znacząco niezależnie od doboru parametrów τ i γ. Wypada on na tle innych najgorzej, ma średnio największy błąd kwadratowy dla każdego przypadku.
Oba estymatory Jamesa-Steina (ściągający do zera — kolor zółty, ściągający do wspólnej średniej — zielony) zachowują się podobnie w każdym z przypadków. Najlepiej wypadają one dla najmniejszych wartości parametrów ($\tau = 1.5\sqrt{2\log 1000},\ \gamma = 0.01$), zwiększenie któregoś z tych parametrów powoduje zwiększenie się błedu. Mają również estymatory Jamesa-Steina zauważalnie największy IQR, co może wpływać na większą wariancję estymatorów.
Z wykresów również możemy zauważyć, że estymatory „ucięte" z zadania 3. mają dużo mniejsze wartości błędów, co zgadza się z analizą przeprowadzoną na podstawie tabeli. Można zauważyć, że ze zwiększającym się parametrem γ błędy estymatorów z zadania 3. się zwiększają, w tym dla procedury Bonferroniego widać największy wzrost.



Rysunek 1: Boxploty błędów kwadratowych dla różnych τ i γ.

Tabela 4: Średnia suma liczby błędów I i II rodzaju przy 1000 powtórzeniach.

γ	τ	BF	BH	Bayess
0.01	$1.5\sqrt{2\log 1000}$	5.15	5.02	5.00
0.05	$1.5\sqrt{2\log 1000}$	2.86	2.93	2.84
0.10	$1.5\sqrt{2\log 1000}$	26.88	23.20	23.22
0.01	$3\sqrt{2\log 1000}$	13.97	12.68	12.31
0.05	$3\sqrt{2\log 1000}$	52.57	42.45	42.32
0.10	$3\sqrt{2\log 1000}$	28.34	24.17	23.67

Ponownie (jak przy jednym powtórzeniu) wyniki są zbliżone, jednak wśród nich najmniejsza liczba błędów, dla każdego z przypadków, jest dla procedury przy użyciu klasyfikatora Bayesowskiego. Dla 1000 powtórzeń również najgorzej wypada procedura Bonferroniego.

Wnioski

- Dla rozpatrywanego modelu regresji liniowej najlepszym estymatorem wektora współczynników okazał się „ucięty" estymator wyznaczony za pomocą klasyfikatora Bayesowskiego. Charakteryzował się on najmniejszym błędem średniokwadratowym niezależnie od doboru parametrów γ oraz τ.

- Najgorszym pod względem MSE okazał się estymator najmniejszych kwadratów.

- Najmniejsza liczba sumy błedów I i II rodzaju dla każdego z przypadków wyszła dla procedury klasyfikatora Bayesowskiego, procedura Benjaminiego-Hochberga wypadła bardzo podobnie.