Analiza dużych zbiorów danych raport nr 2

Dominik Mika

20 kwietnia 2021

W tym raporcie sprawdzimy własności różnych estymatorów parametru β w regresji liniowej.

Dla $p = 1000$ wygenerujemy ortonormalną macierz planu $X_{1000\times1000}$. Następnie wygenerujemy wektor współczynników regresji jako ciąg niezależnych zmiennych losowych z rozkładu

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

gdzie δ_0 jest rozkładem skupionym w 0, a $\phi(0,\tau^2)$ jest gęstością rozkładu normalnego $N(0,\tau^2)$. Rozważymy różne przypadki dla: $\epsilon \in \{0.01, 0.05, 0.1\}$ oraz $\tau \in \{1.5\sqrt{2\log 1000}, 3\sqrt{2\log 1000}\}$. Dla każdego z tych przypadków wygenerujemy wektor odpowiedzi

$$Y = X\beta + \epsilon,$$

gdzie $\epsilon \sim N(0, I_{1000\times1000})$.

W kolejnych zadaniach będziemy wyznaczać różne estymatory oraz obliczymy ich własności. Wyniki przedstawię pod koniec.

Zadanie 1

Estymator najmniejszych kwadratów β^{LS} dla wektora β przedstawia się następująco:

$$\beta^{LS} = (X'X)^{-1}X'Y.$$

Ma on rozkład $\beta^{LS} \sim N(\beta, \sigma^2(X'X)^{-1})$.

W naszym przypadku macierz X była macierzą ortonormalną, więc $X'X = I$. Dodatkowo $\sigma^2 = 1$. Stąd estymator ma postać:

$$\beta^{LS} = X'Y,$$

o rozkładzie $\beta^{LS} \sim N(\beta, I)$.

Zadanie 2

Estymator Jamesa-Steina ściągający do zera przedstawia się następująco:

$$\hat{\beta}^{JS_{zero}} = \left(1 - \frac{p-2}{||\hat{\beta}^{LS}||^2}\right)\hat{\beta}^{LS}$$

Natomiast ściągający do średniej następująco:

$$\hat{\beta}^{JS_{rednia}} = (1-d)\,\hat{\beta}^{LS} + d\overline{\hat{\beta}^{LS}},$$

gdzie $d = \frac{p-1}{p-3}\frac{1}{Var(\hat{\beta}^{LS})}$.

Zadanie 3

i)

Klasyfikator Bayesowski w naszym przypadku działa w następujący sposób:

$$\text{odrzucamy} \quad H_{0i} \quad \text{, gdy} \quad \frac{\phi(\hat{\beta}_i^{LS}, 1+\tau^2)}{\phi(\hat{\beta}_i^{LS}, 1)} > \frac{1-\epsilon}{\epsilon},$$

gdzie ϕ to gęstości rozkładów normalnych o średnich 0 i wariancjach 1 oraz $1+\tau^2$.

ii)

Procedurę Bonferoniego stosujemy w ten sam sposób jak na poprzedniej liście. Oznaczmy p_i jako p-wartość testu dla $i-$tej współrzędnej. Wtedy odrzucamy hipotezę H_{0i}, gdy:

$$p_i \leq \frac{0.05}{1000}.$$

iii)

Procedurę Benjaminiego-Hochberga również stosujemy podobnie jak na poprzedniej liście. Sortujemy wektor p-wartości rosnąco i szukamy największego takiego $i = i_0$, dla którego

$$p_i \leq \frac{i \cdot 0.05}{1000}.$$

Następnie odrzucamy wszystkie H_{0i}, dla których $i \leq i_0$.

Teraz dla każdego z podpunktów **i)**, **ii)** i **iii)** wyznaczymy ucięte estymatory β

$$\hat{\beta}_i = \begin{cases} \hat{\beta}^{LS} & \text{jeżeli odrzucono } H_{0i} : \mu_i = 0 \\ 0 & \text{w przeciwnym wypadku} \end{cases}$$

Jedno powtórzenie

Następnie estymatory z zadań 1-3 porównamy pod kątem błędu średniokwadratowego

$$SE = ||\hat{\beta} - \beta||^2.$$

Dla procedur z zadania 3 wyznaczymy sumę błędów pierwszego i drugiego rodzaju.

Poniżej znajduje się tabelka błędów średniokwadratowych naszych estymatorów.

ϵ	τ	SE_{MLE}	$SE_{JS-\text{zero}}$	$SE_{JS-\text{średnia}}$	SE_{Bayes}	SE_{Bonf}	SE_{BH}
0.01	$1.5\sqrt{2\log 1000}$	1046.6	247.762	247.678	42.315	28.674	42.315
0.05	$1.5\sqrt{2\log 1000}$	971.626	602.131	602.114	51.89	132.403	51.89
0.1	$1.5\sqrt{2\log 1000}$	1107.655	855.345	856.707	381.25	549.58	381.25
0.01	$3\sqrt{2\log 1000}$	1019.58	413.843	413.957	4.452	4.452	4.452
0.05	$3\sqrt{2\log 1000}$	957.511	776.783	776.882	124.858	172.094	141.958
0.1	$3\sqrt{2\log 1000}$	1003.432	959.888	959.651	170.062	264.241	217.765

Na pierwszy rzut oka widzimy, że największe wartości błędu kwadratowego dostajemy dla estymatora największej wiarogodności, czego mogliśmy się spodziewać. Jego wartości są bliskie 1000, co wynika z teoretycznego rachunku:

$$SE(\hat{\beta}^{LS}) = \mathbb{E}||\hat{\beta}^{LS} - \beta||^2 = \sum_{i=1}^{1000} \mathbb{E}[(\hat{\beta}^{LS} - \beta)^2] = 1000\sigma^2 = 1000$$

Zdecydowanie lepiej pod tym kątem wypadły oba estymatory Jamesa-Steina, które osiągały bardzo podobne wartości między sobą. Wynika to z tego, że teoretyczna średnia wynosi zero, więc oba estymatory są ściągające do zera, ale mają nieco inną postać. Dodatkowo widzimy, że błąd średniokwadratowy dla nich dość mocno rośnie wraz ze wzrostem parametrów ϵ i τ.

Ewidentną różnicę widzimy natomiast dla estmatorów Bayesa, Bonferoniego oraz Benjaminiego-Hochberga, ponieważ uzyskujemy dla nich najmniejsze wartości.

Poniżej przedstawiam tabelki zawierające liczba błędów I-go i II-go rodzaju oraz ich sumy.

ϵ	τ	Bayes I-te	Bayes II-te	Bonf I-te	Bonf II-te	B-H I-te	B-H II-te
0.01	$1.5\sqrt{2\log 1000}$	1	8	0	8	1	8
0.05	$1.5\sqrt{2\log 1000}$	0	18	0	25	0	18
0.1	$1.5\sqrt{2\log 1000}$	5	65	0	80	5	65
0.01	$3\sqrt{2\log 1000}$	0	0	0	0	0	0
0.05	$3\sqrt{2\log 1000}$	1	12	0	15	3	11
0.1	$3\sqrt{2\log 1000}$	1	21	0	29	7	20

ϵ	τ	suma Bayes	suma Bonf	suma B-H
0.01	$1.5\sqrt{2\log 1000}$	9	8	9
0.05	$1.5\sqrt{2\log 1000}$	18	25	18
0.1	$1.5\sqrt{2\log 1000}$	70	80	70
0.01	$3\sqrt{2\log 1000}$	0	0	0
0.05	$3\sqrt{2\log 1000}$	13	15	14
0.1	$3\sqrt{2\log 1000}$	22	29	27

Po obserwacji błędów I-go i II-go rodzaju widzimy, że te estymatory w większości przypadków radzą sobie dość dobrze z testowaniem, co sprawdziliśmy na poprzednim raporcie. Widzimy, że wraz ze wzrostem epsilona wzrasta liczba popełnionych błędów. Natomiast dla mniejszego tau popełniliśmy więcej błędów. Wynika to z tego, że wraz ze wzrostem parametru epsilon maleje liczba nieistotnych parametrów β_i. Natomiast dla większych τ wzrasta wariancja β co zwiększa jej zakres wartości i zmniejsza prawdopodobieństwo popełnienia błędu. Z tego samego powodu zmienia się błąd średniokwadratowy, ponieważ ucięte estymatory wyznaczamy właśnie na podstawie wyników testów.

Powtórzenie 1000 razy

Na koniec dla każdej kombinacji ϵ i τ powtórzymy doświadczenie 1000 razy i porównamy analizowane procedury pod kątem $MSE = \mathbb{E}[SE]$ i wartości oczekiwanej sumy liczby błędów pierwszego i drugiego rodzaju.

Poniżej przedstawiają się średnie z wartości dla 1000 powtórzeń.

ϵ	τ	MSE_{MLE}	$MSE_{JS-\text{zero}}$	$MSE_{JS-\text{średnia}}$	MSE_{Bayes}	MSE_{Bonf}	MSE_{BH}
0.01	$1.5\sqrt{2\log 1000}$	999.44	226.41	227.162	31.68	34.849	31.746
0.05	$1.5\sqrt{2\log 1000}$	997.365	602.688	603.097	130.615	176.535	131.2
0.1	$1.5\sqrt{2\log 1000}$	1002.101	754.974	755.205	233.375	351.01	234.78
0.01	$3\sqrt{2\log 1000}$	1000.703	521.006	521.457	24.658	25.746	25.843
0.05	$3\sqrt{2\log 1000}$	1003.035	857.715	857.911	101.994	123.518	105.631
0.1	$3\sqrt{2\log 1000}$	999.289	922.202	922.272	188.448	249.853	192.991

Po powtórzeniu naszego eksperymentu 1000 razy widzimy, że nasze poprzednie obserwacje potwierdzają się w tym przypadku.

ϵ	τ	Bayes I-te	Bayes II-te	Bonf I-te	Bonf II-te	B-H I-te	B-H II-te
0.01	$1.5\sqrt{2\log 1000}$	0.302	4.676	0.049	5.180	0.321	4.663
0.05	$1.5\sqrt{2\log 1000}$	1.731	20.976	0.041	26.405	1.456	21.321
0.1	$1.5\sqrt{2\log 1000}$	3.885	38.425	0.048	52.24	2 2.847	39.707
0.01	$3\sqrt{2\log 1000}$	0.158	2.665	0.04	2.859	0.425	2.47
0.05	$3\sqrt{2\log 1000}$	0.98	11.559	0.05	14.099	2.056	10.806
0.1	$3\sqrt{2\log 1000}$	2.064	21.623	0.039	28.278	3.916	20.263

ϵ	τ	suma Bayes	suma Bonf	suma B-H
0.01	$1.5\sqrt{2\log 1000}$	4.978	5.229	4.984
0.05	$1.5\sqrt{2\log 1000}$	22.707	26.446	22.777
0.1	$1.5\sqrt{2\log 1000}$	42.310	52.288	42.554
0.01	$3\sqrt{2\log 1000}$	2.823	2.899	2.895
0.05	$3\sqrt{2\log 1000}$	12.539	14.149	12.862
0.1	$3\sqrt{2\log 1000}$	23.687	28.317	24.179

Znów możemy zauważyć, że uśrednione wartości mają takie same własności jak te dla jednego powtórzenia. Jedyne co sie zmeniło to dla $\tau = 3\sqrt{2\log 1000}$ średnia liczba błędów zdecydowanie się zmniejszyła.

Podsumowanie

Biorąc pod uwagę nasze wszystkie wyniki mogliśmy zaobserwować różne zachowania. Przede wszystkim wyznaczone ucięte estymatory wektora β otrzymują zdecydowanie mniejsze błędy średniokwadratowe. Dodatkowo ich wartość zależała w przypadku estymatora Bayesa oraz metod Bonferoniego i Benjaminiego-Hochberga od liczby popełnianych błędów.

Obserwowaliśmy, że estymatory Jamesa-Steina są lepsze od estymatora **LS**. Natomiast zdecydowanie lepsze wyniki otrzymywaliśmy dla metod na wielokrotne testowanie. Metoda Bonferoniego dawała nam najgorsze wyniki z tych trzech metod. Najlepiej zachowywał się estymator dla metody Bayesa. Estymator dla metody B-H zachowywał się podobnie, lecz dawał nieco gorsze wyniki.