Raport 3 - Analiza dużych zbiorów danych

Magdalena Potok

2024-05-22

Celem tego raportu jest analiza różnych kryteriów informacyjnych. W tym raporcie skupimy się na kryteriach $AIC, BIC, RIC, mBIC$ oraz $mBIC2$, oceniając je pod kątem liczby prawdziwych odkyć, liczby fałszywych odkryć, błędu średniokwadratowego oraz mocy. Przeanalizujemy również różne estymatory błędu predykcji PE. Rozważymy sytuacje z liczbą obserwacji $n = 1000$ oraz $n = 5000$.

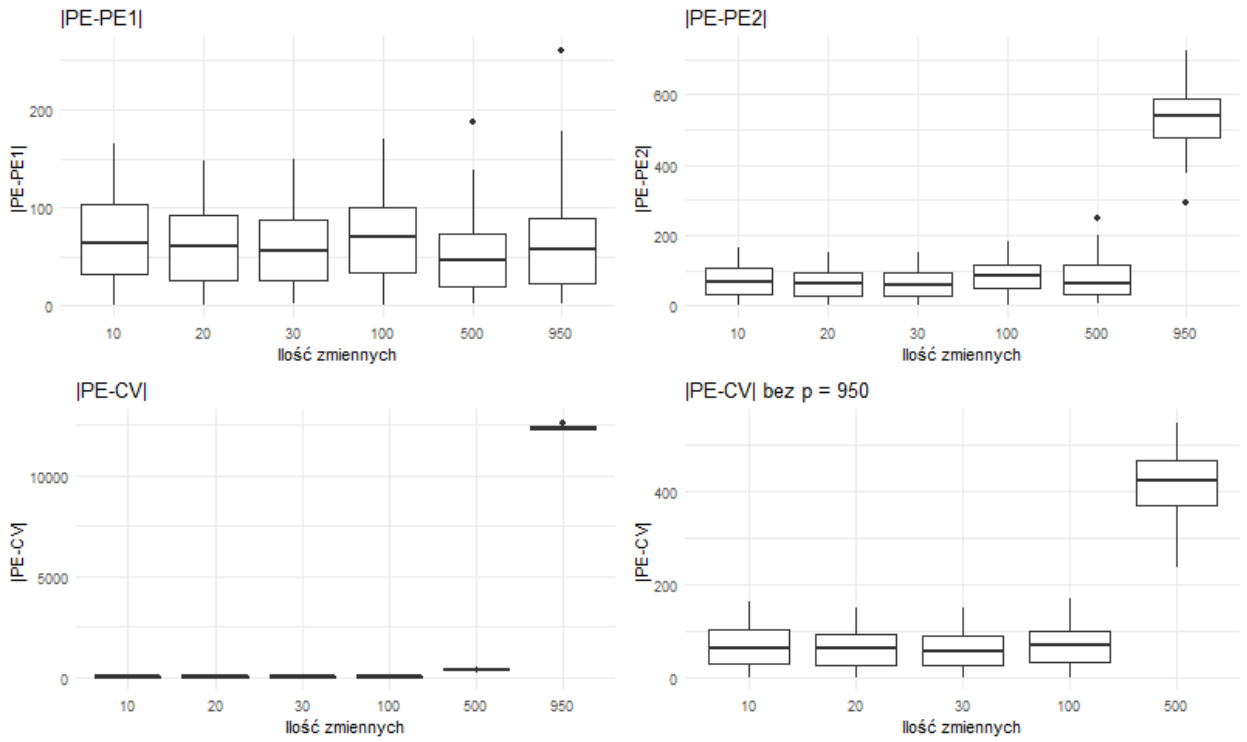
Zadanie 1

Dla macierzy planu $\mathbb{X}_{nx950}$, której elementy są niezależnymi zmiennymi z losowego rozkładu normalnego $N(0, \sigma = \frac{1}{\sqrt{n}})$ wygenerowany został wektor odpowiedzi zgodnie z modelem $Y = \mathbb{X}\beta + \epsilon$, gdzie $\beta = (\beta_1, \ldots, \beta_{950})^T,\ \beta_1 = \ldots \beta_{20} = 3.5,\ i > 20\ \beta_i = 0,\ \epsilon \sim N(0, I)$. Zbudowane zostały modele wykorzystujące $p \in \{10, 20, 50, 100, 200, 500, 950\}$ zmiennych. Dla każdego z modelu wyestymowana została β metodą najmniejszych kwadratów i policzony został błąd predykcji, dwie wersje jego estymatora, RSS oraz CV.

Tabela 1: Obliczone wartości dla pojedyńczej próby.

p	PE	RSS	$\widehat{PE}_1$	$\widehat{PE}_2$	CV
10	1019.12	1146.93	1166.93	1170.10	1170.61
20	1080.42	1042.79	1082.79	1085.35	1086.17
30	1006.35	1039.03	1099.03	1103.30	1105.18
50	1067.10	1008.70	1108.70	1114.88	1119.12
100	1114.14	964.25	1164.25	1178.53	1191.95
500	1575.14	531.90	1531.90	1595.71	2142.27
950	2035.21	62.08	1962.08	2421.17	26436.47

Z tabeli możemy odczytać, że model zbudowany na modelu wykorzystującym 20 pierwszych kolumn oryginalnej macierzy planu okazuje się najlepszy pod wzlędem wszystkich estymatorów PE. Dla $p = 30$ mamy zbliżone wartości, ten model również wypada dobrze. Wraz ze wzrostem p obliczona wartość PE oraz estymatory PE wzrastają, dla mniejszego $p = 10$ również są większe. Możemy również odczytać, że RSS nie jest dobrą miarą jakości modelu, wraz z rozszerzaniem macierzy planu, nawet o nieistotne zmienne, ta wartość maleje.
Powyższy eksperyment został powtórzony 100 razy i zostały utworzone wykresy pudełkowe wartości bezwzględnej różnicy między prawdziwą wartością statystyki PE, a jej estymatorem (Rysunek 1.). Dzięki wykresom możemy zauważyć jak słabo estymator CV radzi sobie dla dużej liczby kolumn macierzy planu. Dla $p = 950$ $|PE - CV|$ osiąga bardzo duże wartości. Można zauważyć, że dla mniejszej liczby zmiennych estymator $\widehat{PE}_2$ oraz CV wypadają podobnie, jak $\widehat{PE}_1$, jednak dla $p = 500$ $\widehat{PE}_2$ wypada nieco gorzej, a CV osiąga już duże wartości. Dla $p = 950$ CV oraz $\widehat{PE}_2$ wypadają znacznie gorzej niż $\widehat{PE}_1$.



Rysunek 1: Wykresy pudełkowe wartości bezwzględnej różnicy między PE a estymatorem.

Zadanie 2

Zastosujemy *AIC, BIC, RIC, mBIC* i *mBIC2* do identyfikacji istotnych zmiennych w bazach danych składających się z $p \in \{50, 100, 200, 500, 950\}$ zmiennych. Policzone zostały liczby prawdziwych i fałszywych odkryć oraz błąd estymacji wektora $\mathbb{E}Y = X\beta$.

Tabela 2: Obliczone wartości dla różnych kryteriów informacyjnych.

	p = 50					p = 100				
	TD	FD	Power	FDR	SE	TD	FD	Power	FDR	SE
AIC	20	5	1.00	0.2	31.746	20	17	1.00	0.459	72.431
BIC	16	0	0.80	0.0	57.632	16	0	0.80	0.000	57.632
RIC	17	0	0.85	0.0	45.133	15	0	0.75	0.000	70.921
mBIC	10	0	0.50	0.0	131.463	9	0	0.45	0.000	147.791
mBIC2	14	0	0.70	0.0	81.765	14	0	0.70	0.000	81.765

	p = 500					p = 950				
	TD	FD	Power	FDR	SE	TD	FD	Power	FDR	SE
AIC	20	50	1.0	0.714	192.868	20	50	1.00	0.714	192.868
BIC	16	3	0.8	0.158	87.168	16	11	0.80	0.407	144.075
RIC	10	1	0.5	0.091	143.993	9	0	0.45	0.000	147.791
mBIC	6	0	0.3	0.000	174.291	5	0	0.25	0.000	192.576
mBIC2	10	1	0.5	0.091	143.993	5	0	0.25	0.000	192.576

W tabeli zostały pominięte wyniki dla $p = 200$, są one bardzo zbliżone do wartości statystyk dla $p = 500$. Liczba prawdziwych odkryć jest w każdym z przypadków maksymalna ($= 20$) dla kryterium AIC, w przypadku

kryterium BIC również jest to liczba stała dla każdego p, ale nie został osiągnięty maksymalny wynik. Pozostałe kryteria: $RIC, mBIC, mBIC2$ wraz ze wzrostem liczby zmiennych objaśniających posiadają coraz mniejszą liczbę prawdziwych odkryć. Najgorzej wśród nich wypada $mBIC$.
W porównaniu do pozostałych kryteriów AIC ma bardzo dużą liczbę fałszywych odkryć dla każdego przypadku. Ze wzrostem p obliczona wartość FD rośnie dla kryterium BIC, ale nie jest to tak gwałtowy wzrost, jak w przypadku AIC. Dla pozostałych kryteriów liczba fałszywych odkryć dla każdej liczby p jest bardzo mała. Ze zwiększającą się liczbą regresorów można zauważyć, że dla każdego kryterium zwiększają się wartości SE. Najgorzej wśród wszystkich kryteriów wypada $mBIC2$ - ma on największe SE za każdym razem. Dla mniejszej liczby p najlepiej wypadają AIC, RIC oraz BIC, dla większych p zauważalnie najlepiej wypada kryterium BIC.

Powyższy eksperyment został powtórzony 100 razy, wyniki w tabeli przedstawiają uśrednioną moc, estymowane FDR oraz estymowane MSE.

Tabela 3: Uśrednione wartości dla różnych kryteriów informacyjnych przy 100 krotnym powtórzeniu.

	p = 50			p = 100			p = 200		
	Power	FDR	MSE	Power	FDR	MSE	Power	FDR	MSE
AIC	0.95	0.240	53.120	0.95	0.367	68.466	0.95	0.558	121.825
BIC	0.60	0.000	100.461	0.60	0.000	100.461	0.60	0.077	107.250
RIC	0.65	0.071	95.513	0.40	0.000	150.663	0.30	0.000	175.254
mBIC	0.25	0.000	184.363	0.20	0.000	192.561	0.20	0.000	192.561
mBIC2	0.40	0.000	150.663	0.25	0.000	184.363	0.20	0.000	192.561

	p = 500			p = 950		
	Power	FDR	MSE	Power	FDR	MSE
AIC	0.95	0.729	207.468	0.95	0.729	207.468
BIC	0.60	0.143	111.211	0.60	0.294	129.660
RIC	0.20	0.000	192.561	0.20	0.000	192.561
mBIC	0.10	0.000	212.293	0.10	0.000	212.293
mBIC2	0.10	0.000	212.293	0.10	0.000	212.293

Tak jak w przypadku, gdy eksperyment został powtórzony tylko raz, największą moc otrzymujemy przy zastosowaniu kryterium AIC, a nieco gorzej wypada BIC. Dla pozostałych kryteriów, wraz ze wzrostem liczby regresorów, zmniejsza się ta statystyka, najgorzej wśród nich wypadają kryteria $mBIC$ oraz $mBIC2$. Jeżeli chodzi o wartość statystyki FDR najgorzej wypada kryterium AIC, dla niego ta wartość jest największa wśród innych kryteriów oraz rośnie wraz ze wzrostem liczby regresorów. Dla pozostałych kryteriów wartość FDR jest bliska 0, w przypadku BIC rośnie ze wzrostem regresorów, ale nie tak gwałtownie, jak AIC.
W przypadku wartości statystyki MSE najgorzej wypada kryterium $mBIC$, a ze zwiększającą się liczbą zmiennych objaśniających $mBIC2$ zbliża się do tych samych wyników. Dla małych wartości p najlepiej radzi sobie AIC, ze wzrostem p lepiej radzi sobie BIC.

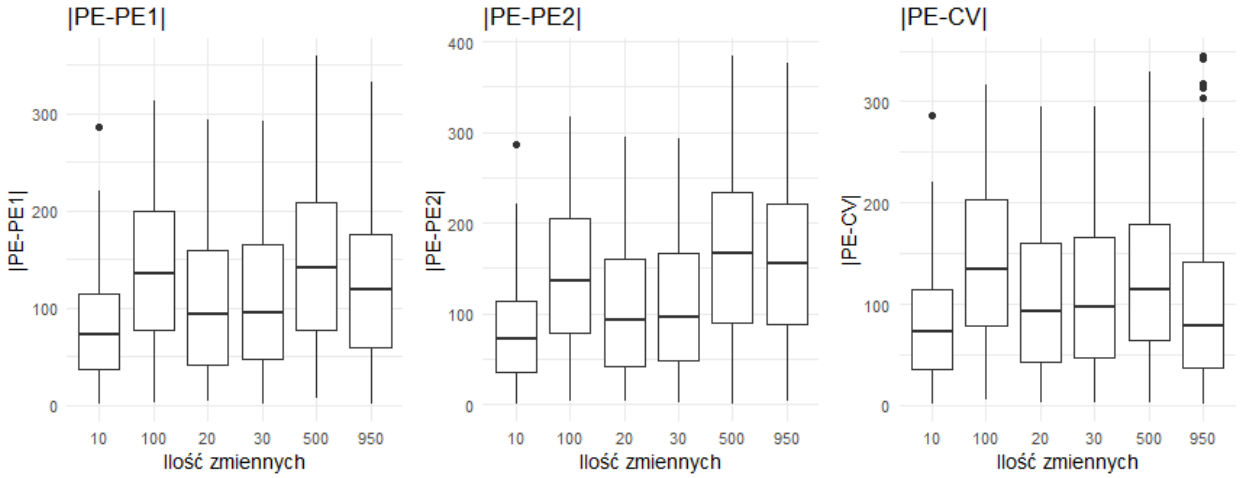
Zadanie 3

Powyższe zadania zostaną powtórzone w sytuacji, gdy $n = 5000$. Dla obliczeń związanych z kryteriami informacyjnymi pominięta została liczba regresorów $p = 50$, ponieważ ten przypadek generował błąd.

Tabela 4: Obliczone wartości dla pojedyńczej próby, $n = 5000$.

p	PE	RSS	$\widehat{PE}_1$	$\widehat{PE}_2$	CV
10	4960.90	5096.65	5116.65	5117.08	5117.02
20	5017.21	4977.57	5017.57	5017.55	5017.28
30	5053.33	4963.11	5023.11	5023.03	5023.04
50	5075.91	4946.38	5046.38	5046.31	5046.88
100	5004.51	4909.79	5109.79	5110.19	5112.50
500	5728.79	4494.43	5494.43	5493.20	5550.86
950	6075.69	4042.49	5942.49	5938.97	6164.13

Prawie wszystkie wyniki są znacząco większe niż, gdy $n = 1000$, dzieję się tak, ponieważ p jest zacząco mniejsze od n. Estymatory najbliżej wartości statystyki PE zostały policzone dla modelu zbudowanego na 10 regresorach, ten model ma najmniejszą wartość PE. Można z tabeli zauważyć, że nie ma żadnej znacząco odbiegającej wartości, tak jak dla przypadku $n = 1000$.



Rysunek 2: Wykresy pudełkowe wartości bezwzględnej różnicy między PE a estymatorem, $n = 5000$.

Z wykresów pudełkowych możemy odczytac, że mniejwięcej wszystkie estymatory PE zachowują się dość podobnie. Model z najbliższymi do PE wartościami estymatorów, to model zawierający 10 zmiennych objaśniających. Zgadza się ta obserwacja z wynikami wyliczonymi w Tabeli 4.

Tabela 5: Obliczone wartości dla różnych kryteriów informacyjnych, $n = 5000$.

	p = 100					p = 200				
	TD	FD	Power	FDR	SE	TD	FD	Power	FDR	SE
AIC	20	14	1.00	0.412	57.843	20	28	1.00	0.583	104.359
BIC	14	0	0.70	0.000	81.277	14	0	0.70	0.000	81.277
RIC	16	0	0.80	0.000	55.844	13	0	0.65	0.000	93.363
mBIC	8	0	0.40	0.000	155.071	5	0	0.25	0.000	190.444
mBIC2	9	0	0.45	0.000	143.069	9	0	0.45	0.000	143.069

	p = 500					p = 950				
	TD	FD	Power	FDR	SE	TD	FD	Power	FDR	SE
AIC	20	50	1.00	0.714	186.568	20	50	1.00	0.714	186.568
BIC	14	2	0.70	0.125	100.364	14	5	0.70	0.263	129.584
RIC	11	0	0.55	0.000	117.177	9	0	0.45	0.000	143.069
mBIC	3	0	0.15	0.000	216.450	2	0	0.10	0.000	225.727
mBIC2	8	0	0.40	0.000	155.071	5	0	0.25	0.000	190.444

Wyniki są zbliżone do wyników z tabeli 2. dla $n = 1000$. Kryterium AIC osiąga w każdym z przypadków maksymalną liczbę prawdziwych odkryć, ale za każdym razem osiąga też największą liczbę fałszywych odkryć. Ze wzrostem liczby regresorów liczba fałszywych odkryć kryteriów AIC i BIC rośnie. Najmniejsza liczba prawdziwych odkryć dla każdego p wychodzi dla kryterium $mBIC$, gdzie $mBIC2$ wypada niewiele lepiej. Pod względem FDR ponownie najgorzej wychodzi kryterium AIC, a przy większej liczbie zmiennych objaśniających ta statystyka dla kryterium BIC również rośnie. Kryterium $mBIC$ wypada również najgorzej pod względem statystyki SE. Wnioski z tabeli 2. i tabeli 5. są identyczne.

Tabela 6: Uśrednione wartości dla różnych kryteriów informacyjnych przy 100 krotnym powtórzeniu, $n = 5000$.

	p = 100			p = 200			p = 950		
	Power	FDR	MSE	Power	FDR	MSE	Power	FDR	MSE
AIC	1.00	0.333	59.880	1.00	0.565	119.663	1.00	0.714	198.867
BIC	0.80	0.059	77.546	0.80	0.059	77.546	0.80	0.200	109.507
RIC	0.85	0.056	66.087	0.75	0.000	82.386	0.60	0.000	118.602
mBIC	0.55	0.000	131.231	0.35	0.000	177.337	0.25	0.000	199.774
mBIC2	0.70	0.067	102.375	0.70	0.000	94.303	0.35	0.000	177.337

W tabeli pominięte zostały wyniki dla $p = 500$, ponieważ zbliżone były do wyników dla $p = 950$.
Ponownie, jak przy jednokrotnym eksperymencie, największą moc otrzymujemy przy zastosowaniu kryterium AIC, ale również wtedy otrzymujemy najwiekszą wartość statystyki FDR. Najgorzej pod względem MSE wypada krterium $mBIC$. Ponownie wnioski są podobne, jak do tabeli 3, gdzie $n = 1000$.

Podsumowanie

- Dla mniejszej liczby obserwacji estymatory $\widehat{PE}_1$ oraz $\widehat{PE}_2$ przy rosnącej liczbie regresorów radzą sobie znacznie lepiej niż CV pod względem wartości bezwzględnej między prawdziwą watością PE a estymatorem. Dla większej liczby obserwacji wszystkie estymatory radzą sobie podobnie.

- Z rosnącą liczbą regresorów obliczone wartości FDR rosną dla kryteriów AIC i BIC dla $n = 5000$ oraz $n = 1000$. Najlepiej z tą statystyką radzi sobie kryterium $mBIC$.

- W przypadku kryteriów $RIC, mBIC$ i $mBIC2$ ich moc maleje ze wzrostem liczby zmiennych objaśniających i osiąga bardzo małe wartości dla $n = 5000$ oraz $n = 1000$. Kryterium AIC zawsze ma maksymalną moc albo moc bliską wartości 1.

- Nie ma znaczącej różnicy we wnioskach dotyczących kryteriów informacyjnych między $n = 5000$, a $n = 1000$.