

Analiza dużych zbiorów danych raport nr 3

Dominik Mika

10 maja 2021

W tym raporcie do poniższych analiz będziemy korzystali z wygenerowanych przez nas danych. Najpierw wygenerujemy macierz planu $X_{n \times 950}$, dla $n = 1000$ oraz $n = 5000$ (wyniki dla $n = 5000$ przedstawię w zadaniach 1 i 2), tak że jej elementami są niezależne zmienne losowe z rozkładu $N(0, \frac{1}{\sqrt{1000}})$. Następnie wygenerujemy wektor odpowiedzi Y zgodnie ze wzorem

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

gdzie $\beta_1 = \dots = \beta_{20} = 3.5$, $\beta_{21} = \dots = \beta_{950} = 0$, a $\epsilon \sim N(0, I)$.

Zadanie 1

To zadanie wykonamy dla modeli w oparciu o pierwsze: 10, 20, 30, 100, 500 i 950 zmiennych.

a) Dla każdego z modeli wyestymujemy wektor β metodą najmniejszych kwadratów i wyznaczymy $RSS = \|\hat{Y} - Y\|^2$ oraz wyliczymy wartość oczekiwaną błędu predykcji:

$$PE = \mathbb{E}_{\epsilon^*} \|X(\beta - \hat{\beta}) + \epsilon^*\|^2,$$

gdzie $\epsilon^* \sim N(0, I)$ jest wektorem niezależnym od próby treningowej.

Teoretyczne PE wynosi:

$$PE = \mathbb{E}_{\epsilon^*} \|\tilde{X}\beta - X\hat{\beta} + \epsilon^*\|^2 = \|(I - H)\tilde{X}\beta\|^2 + \sigma^2(n + p),$$

gdzie $H = X(X'X)^{-1}X'$, p - liczba zmiennych w modelu, a \tilde{X} to rzeczywista macierz z której pochodzą dane.

n=1000

W tej części przedstawię wyniki dla $n = 1000$.

Po wyliczeniu tej wartości teoretycznej wyniki PE przedstawiają się w poniższej tabeli:

Il. zmiennych	10	20	30	100	500	950
PE	1130.04	1020	1030	1100	1500	1950

Widzimy, że dla wszystkich przypadków oprócz modelu dla 10 zmiennych PE wynosi $\sigma^2(n + p)$. Oznacza to, że składnik $\|(I - H)X\beta\|^2$ jest bardzo bliski zera. Natomiast w modelu dla 10 zmiennych jest on niezerowy. Wynika to z tego, że w tym przypadku bierzemy okrojaną macierz planu bez 10 istotnych kolumn, czyli zakładamy, że dane pochodzą z innej macierzy niż w rzeczywistości. Stąd ten składnik stanowi znaczący wkład w PE .

b) Następnie używając RSS wyestymujemy PE wykorzystując prawdziwą wartość σ i zastępując ją jej klasycznym nieobciążonym estymatorem.

c) Skorzystamy jeszcze z innej metody estymacji PE , mianowicie stosując walidację krzyżową typu „leave-one-out”.

W celu estymacji korzystamy z trzech następujących wzorów:

1. $PE_t = RSS + 2\sigma^2p$ - znane σ ,
2. $PE_e = RSS + 2p\frac{RSS}{n-p}$ - estymowana σ ,
3. $CV = \sum_{i=1}^n (Y_i - \hat{Y}[i])^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - H_{[i,i]}} \right)^2$ - walidacja krzyżowa.

Wyniki dla tych metod prezentują się w poniższej tabeli:

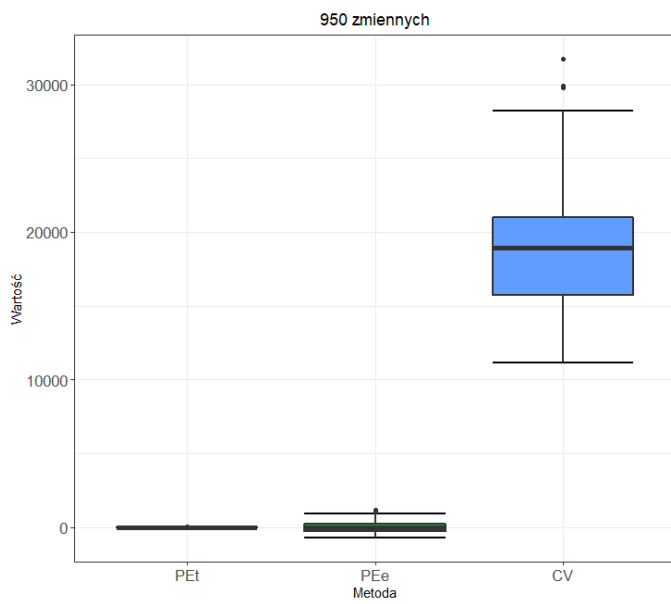
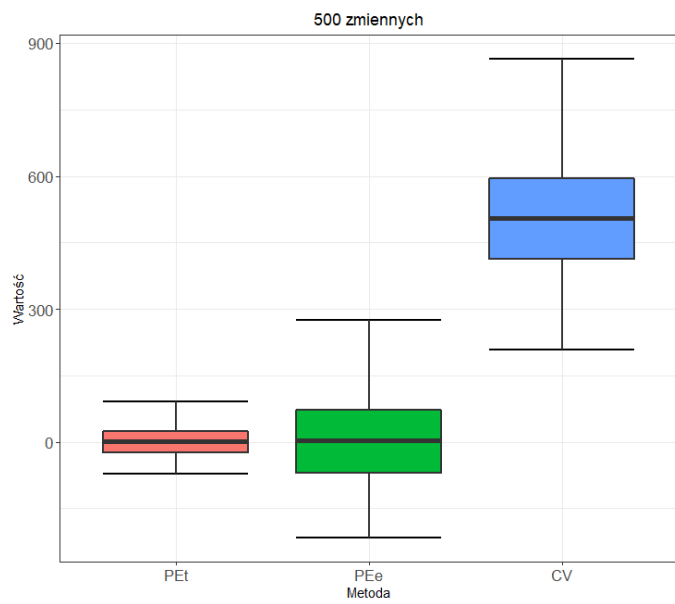
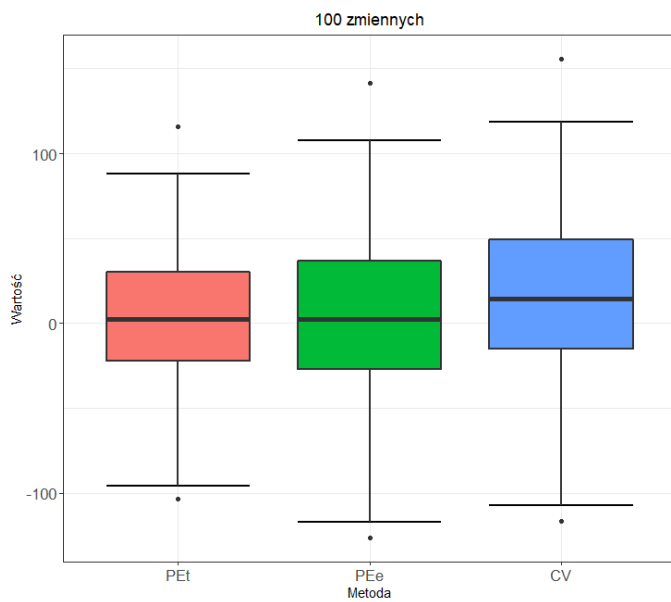
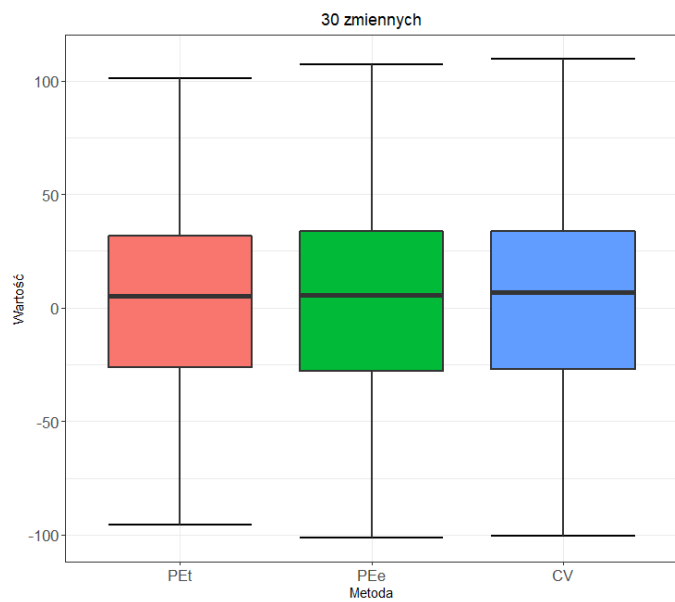
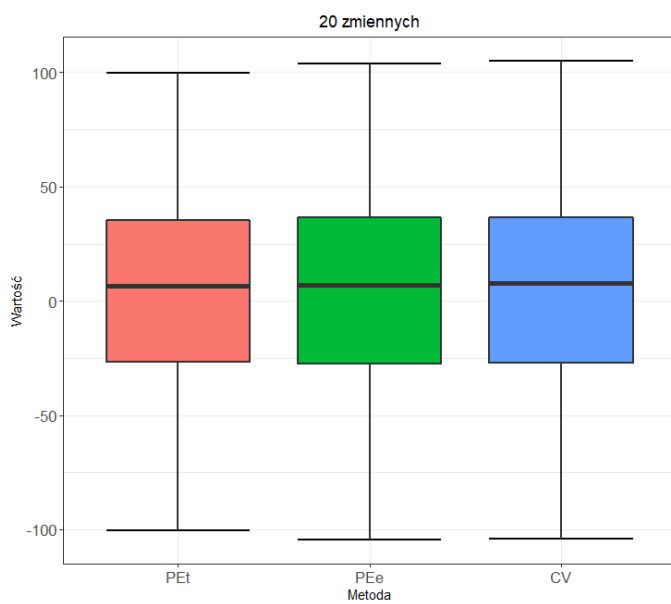
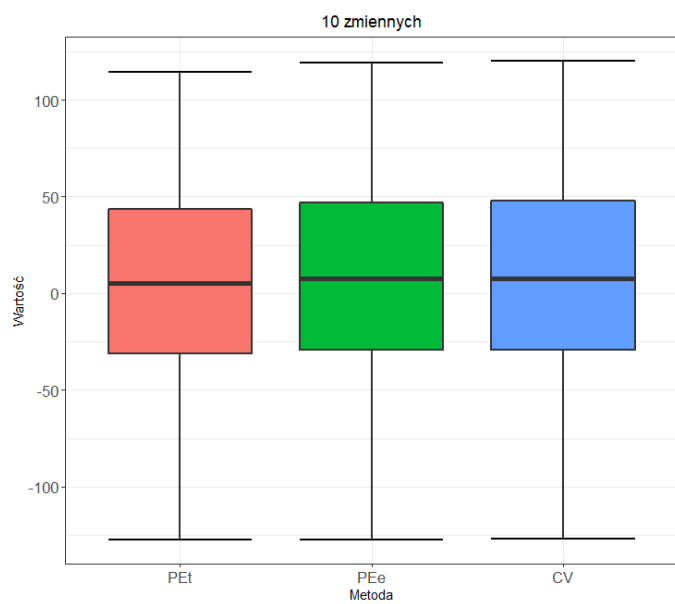
	10	20	30	100	500	950
RSS	1063.30	948.33	937.96	857.50	465.44	46.90
PE_t	1083.30	988.33	997.96	1057.50	1465.44	1946.90
PE_e	1084.78	987.04	995.98	1048.06	1396.33	1829.12
CV	1084.43	986.74	996.19	1059.12	1863.05	19613.44

Widzimy, że dla pierwszych trzech modeli wszystkie metody dość dobrze estymują PE . Wyniki nie różnią się od teoretycznej wartości o więcej niż 60. Natomiast różnicę możemy zauważyć dla pozostałych modeli z 500 i 950 zmiennymi. Widzimy wtedy, że najlepiej estymuje teoretyczną wartość PE_t . Nieco gorzej natomiast PE_e , wynika to stąd, że w pierwszym przypadku zakładamy znajomość σ , a w drugim ją estymujemy, co może zachwiać nasze wyniki. Natomiast walidacja krzyżowa w tych przypadkach nie radziła sobie kompletnie. Wyniki są dziesięciokrotnie większe od tych teoretycznych. Na tej podstawie możemy stwierdzić, że najlepszym estymatorem PE jest PE_t .

Na podstawie estymatorów PE wybralibyśmy model dla 20 zmiennych.

d) Następnie powtórzmy powyższe analizy 100 razy i dla modeli porównamy wykresy pudełkowe wartości $\hat{PE} - PE$ dla trzech wyżej wymienionych estymatorów PE .

Wykresy pudełkowe przedstawiam poniżej:



Na wykresach pudełkowych możemy zaobserwować podobne zjawiska jak te, które widzieliśmy w poprzedniej tabelce. Pierwsze cztery wykresy nie wykazują znaczącej różnicy między tymi metodami. Lecz na dwóch ostatnich widzimy, że skuteczność CV znacznie spada. Natomiast metoda PE_t wydaje się być najlepsza, a metoda PE_e nieco gorsza.

Podsumowując wyniki możemy powiedzieć, że najlepszą metodą do estymacji PE jest metoda PE_t . Natomiast walidacja krzyżowa nie radzi sobie, gdy liczba kolumn w macierzy planu p zaczyna zbliżać się do liczby wierszy n .

n=5000

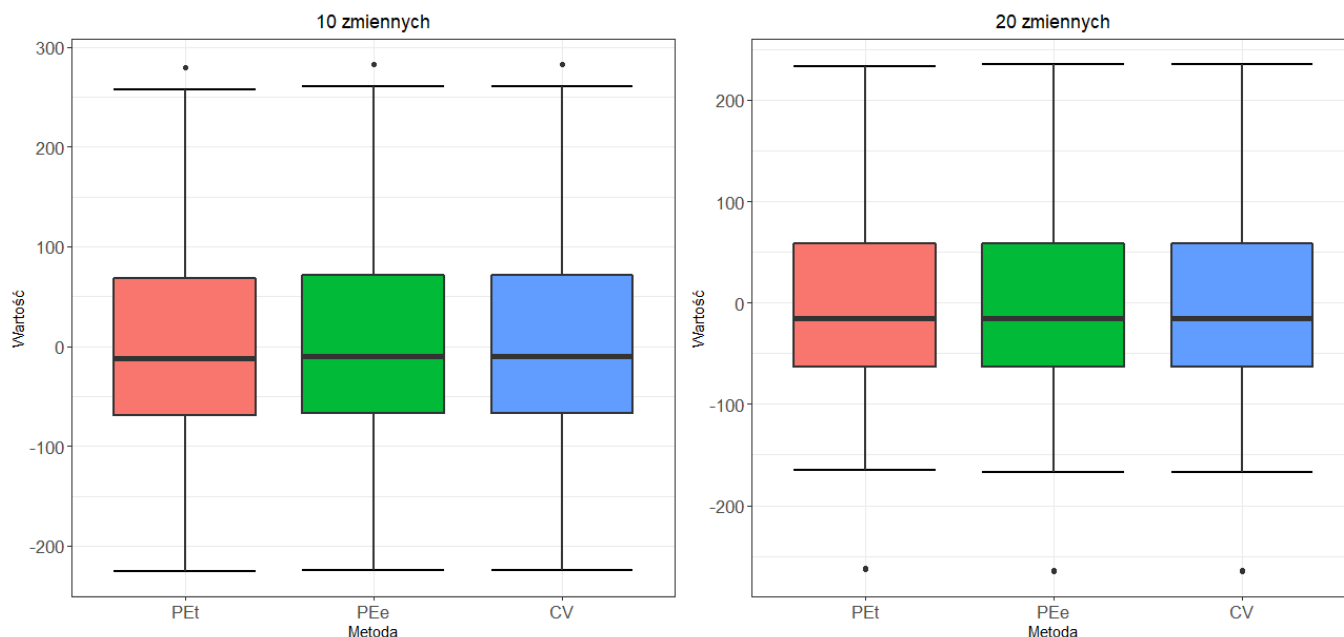
W tej części przedstawię wyniki dla $n = 5000$.

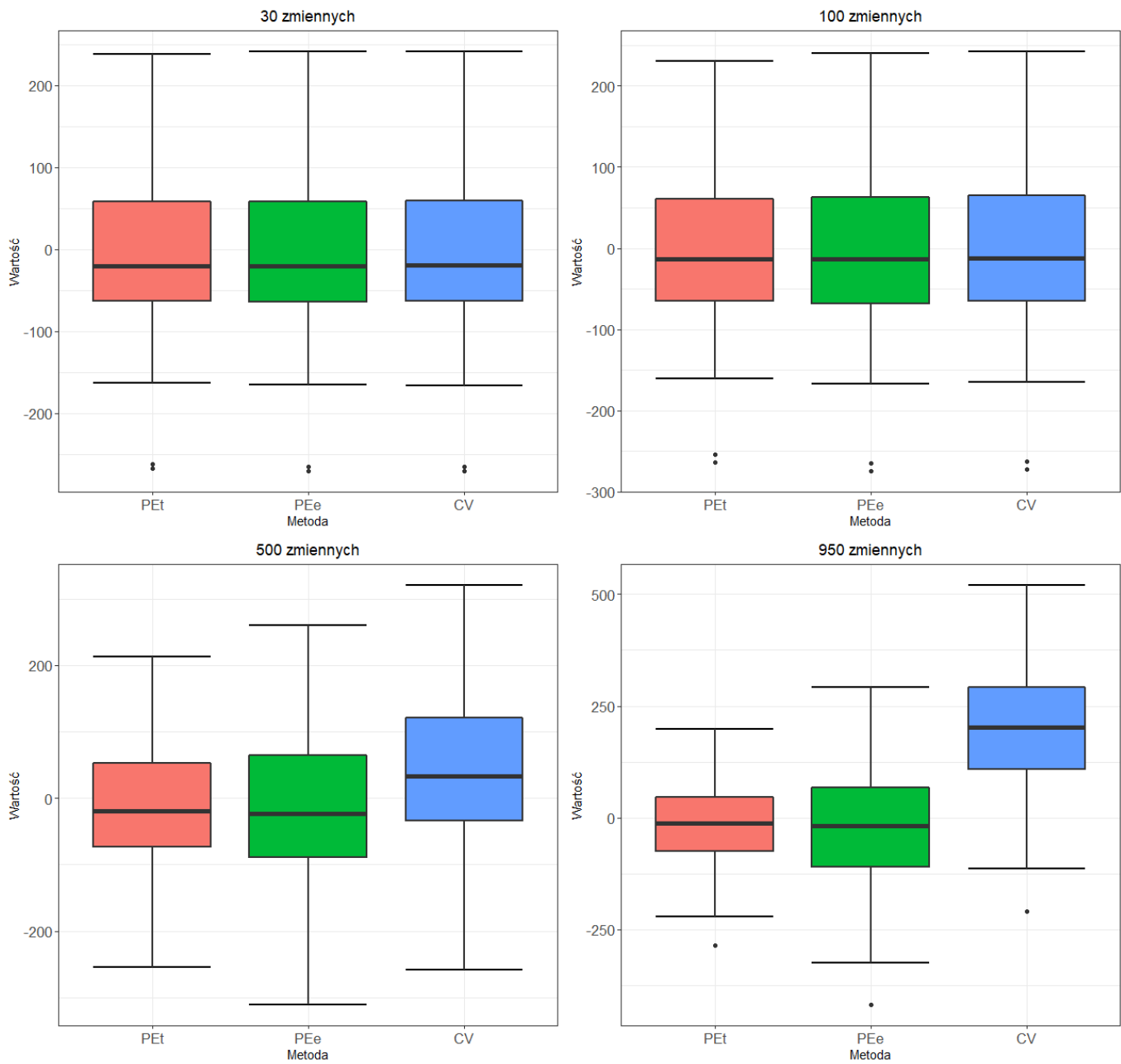
W poniższej tabeli znajdują się teoretyczne wartości PE oraz wartości jego estymatorów.

	10	20	30	100	500	950
PE	5609.94	5020	5030	5100	5500	5950
RSS	5706.56	5070.25	5062.51	5006.07	4586.17	4103.36
PE_t	5726.56	5110.25	5122.51	5206.07	5586.17	6003.36
PE_e	5729.43	5110.97	5123.63	5210.40	5605.32	6028.39
CV	5729.34	5110.70	5123.67	5212.11	5662.19	6260.49

W przypadku, gdy $n = 5000$ widzimy znaczącą różnicę co do przypadku, gdy $n = 1000$. Pierwsze co rzuca się w oczy, to ogromna poprawa CV . Możemy powiedzieć, że ta metoda faktycznie estymuje PE dla podanych przypadków. Natomiast znów najlepszym estymatorem PE w tym przypadku jest PE_t .

Spójrzmy jeszcze na wykresy pudełkowe wartości $\hat{PE} - PE$, dla różnych metod estymacji PE .





W tym przypadku również obserwujemy praktycznie te same zjawiska co w poprzedniej tabeli. Dla pierwszych czterech modeli metody są niemal takie same. Dopiero na ostatnich dwóch widzimy, że CV jest nieco gorsze, a PE_t nieco lepsze.

Dodatkowo, w tym przypadku również na podstawie estymatorów wybralibyśmy model z pierwszymi 20 zmiennymi.

Podsumowując całe zadanie możemy powiedzieć, że najlepszym estymatorem PE jest PE_t , lecz PE_e też radziło sobie dobrze, ale z wspomnianych przeze mnie względów oczywiście nieco gorzej. Natomiast metoda CV dawała nam różne wyniki. Możemy powiedzieć, że radzi ona sobie dobrze, gdy n jest zdecydowanie większe, niż p .

Zadanie 2

W tym zadaniu użyjemy kryteriów BIC, AIC, RIC, mBIC i mBIC2 do identyfikacji istotnych zmiennych w bazach danych składających się z pierwszych 50, 200, 500 i 950 zmiennych. Następnie wyznaczmy następujące wartości.

a) Podamy liczbę prawdziwych i fałszywych odkryć i kwadratowy błąd estymacji wektora $\mathbb{E}Y = X\beta$:

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

n=1000

Najpierw przedstawię wyniki dla $n = 1000$.

<i>SE</i>					
	AIC	BIC	RIC	mBIC	mBIC2
50	38.186	86.079	95.794	138.135	58.204
200	124.546	86.079	126.455	182.533	126.455
500	178.417	119.242	138.135	223.762	150.118
950	178.417	139.172	164.456	223.762	207.075

Widzimy, że wartości *SE* nie różnią się bardzo od siebie dla różnych kryteriów. Dla 50 zmiennych najmniejsze wartości dostajemy dla AIC. Natomiast podsumowując wszystkie, to najmniejsze *SE* dostajemy dla kryterium BIC, a największe dla mBIC.

Fałszywe odkrycia						Prawdziwe odkrycia					
	AIC	BIC	RIC	mBIC	mBIC2		AIC	BIC	RIC	mBIC	mBIC2
50	5	1	1	0	2	50	20	14	13	9	17
200	33	1	0	0	0	200	20	14	10	5	10
500	50	8	0	0	1	500	20	16	9	2	9
950	50	11	0	0	0	950	20	17	7	2	3

Patrząc na fałszywe odkrycia widzimy, że kryterium AIC popełnia ich bardzo dużo. Kryterium BIC natomiast już mniej, a pozostałe trzy metody praktycznie w ogóle. Widzimy, że dla AIC i BIC liczba fałszywych odkryć rośnie wraz z ilością zmiennych w modelu, co wydaje się być zrozumiałe.

Natomiast co do prawdziwych odkryć, widzimy, że kryterium AIC wykryło wszystkie istotne zmienne, kryterium BIC około 15/20, kryterium RIC i BIC nieco mniej, a kryterium mBIC mało.

Próbując znaleźć najoptymalniejsze kryterium moglibyśmy wybrać kryterium RIC lub mBIC2, ponieważ nie popełniają one prawie w ogóle fałszywych odkryć i wykrywają mniej więcej 50% istotnych zmiennych.

b) Następnie powtórzmy punkt a) 100 razy i podamy wyestymowaną moc, FDR i średni błąd kwadratowy estymacji $\mathbb{E}Y = X\beta$ dla wszystkich powyższych kryteriów.

Poniżej znajduje się tabela z wartościami MSE:

<i>MSE</i>					
	AIC	BIC	RIC	mBIC	mBIC2
50	40.75	67.48	78.55	139.78	58.70
200	124.88	76.68	121.84	182.28	111.10
500	189.87	105.40	150.83	199.50	150.91
950	189.87	148.98	168.15	209.55	179.31

Widzimy, że uśrednione wyniki pokazują nam to samo co te w przypadku *SE*.

Natomiast tutaj mamy wartości FDR oraz mocy:

Moc						FDR					
	AIC	BIC	RIC	mBIC	mBIC2		AIC	BIC	RIC	mBIC	mBIC2
50	0.97	0.77	0.72	0.45	0.81	50	0.19	0.02	0.01	0.01	0.02
200	0.97	0.78	0.53	0.27	0.58	200	0.60	0.09	0.02	0.00	0.03
500	0.97	0.78	0.41	0.20	0.41	500	0.72	0.25	0.03	0.00	0.03
950	0.97	0.77	0.33	0.15	0.29	950	0.72	0.41	0.02	0.00	0.01

Tutaj również obserwujemy podobne zjawiska co dla prawdziwych i fałszywych odkryć. Kryterium AIC ma bardzo dużą moc. Kryteria BIC, RIC, mBIC2 już nieco mniejszą, a kryterium mBIC najmniejszą, ale ma ono najmniejsze FDR. Możemy, zauważyć tutaj pewną zależność między mocą i FDR. Kryteria mające większą moc mają również większe FDR, a kryteria z mniejszą mocą mają mniejsze FDR.

Stąd w tym przypadku ciężko wybrać najlepsze kryterium. Kryterium AIC ma bardzo dużą moc, ale kosztem FDR. Moglibyśmy go stosować, gdy chcielibyśmy na pewno wykryć wszystkie istotne zmienne nie bojąc się fałszywych odkryć. Natomiast kryteria BIC, RIC i mBIC2 są bardziej uniwersalne, ponieważ za równo ich moc nie jest mała jak i FDR nie jest duże. Zostało nam jeszcze kryterium mBIC, które ma małą moc, ale z drugiej strony prawie zerowe FDR. Z niego moglibyśmy korzystać, gdybyśmy chcieli ostrożnie dobierać zmienne do naszego modelu obawiając się fałszywego odkrycia.

n=5000

Teraz przedstawię wyniki dla $n = 5000$.

a) Poniżej znajduje się wartość SE oraz fałszywe i prawdziwe odkrycia, dla różnych kryteriów:

SE					
	AIC	BIC	RIC	mBIC	mBIC2
50	49.73	33.54	33.54	33.54	33.54
200	116.94	42.77	33.54	33.54	33.54
500	222.40	84.52	65.70	52.20	65.70
950	222.40	125.65	52.20	33.54	65.70

Widzimy, że dla $n = 5000$, nasze wyniki stają się zupełnie inne. Kryterium AIC ma zdecydowanie największe SE , kryterium BIC drugie największe, a pozostałe trzy są porównywalne.

Fałszywe odkrycia						Prawdziwe odkrycia					
	AIC	BIC	RIC	mBIC	mBIC2		AIC	BIC	RIC	mBIC	mBIC2
50	4	0	0	0	0	50	20	20	20	20	20
200	23	1	0	0	0	200	20	20	20	20	20
500	50	4	2	1	2	500	20	20	20	20	20
950	50	8	1	0	2	950	20	20	20	20	20

Również tutaj możemy zauważyć zdecydowaną zmianę. Choć liczba fałszywych odkryć nie zmieniła się od przypadku dla $n = 1000$, oprócz tych dla kryterium BIC. Natomiast widzimy, że teraz wszystkie kryteria odkrywają wszystkie istotne zmienne.

b) Poniżej przedstawiam uśrednione wyniki:

MSE					
	AIC	BIC	RIC	mBIC	mBIC2
50	37.74	21.80	22.29	21.33	22.67
200	121.77	27.78	23.80	21.50	24.12
500	198.61	40.32	24.22	21.77	24.34
950	198.61	57.50	24.02	21.40	24.71

Wartości MSE przedstawiają się podobnie do poprzednich z SE . Natomiast teraz możemy zobaczyć, że najmniejsze wyniki dostajemy dla kryterium mBIC.

Moc						FDR					
	AIC	BIC	RIC	mBIC	mBIC2		AIC	BIC	RIC	mBIC	mBIC2
50	1.00	1.00	1.00	1.00	1.00	50	0.187	0.003	0.006	0.000	0.008
200	1.00	1.00	1.00	1.00	1.00	200	0.579	0.030	0.010	0.001	0.011
500	1.00	1.00	1.00	1.00	1.00	500	0.714	0.082	0.009	0.001	0.010
950	1.00	1.00	1.00	1.00	1.00	950	0.714	0.146	0.008	0.000	0.010

Widzimy, że wszystkie kryteria mają moc 1. Stąd najlepsze kryterium możemy wybrać patrząc na FDR. Widzimy, że AIC ma dość duże FDR, BIC na akceptowalnym poziomie, RIC i mBIC2 bardzo niskie. Natomiast kryterium mBIC bliskie 0. Zatem zdecydowanie najlepszym kryterium w tym przypadku jest mBIC, niezależnie od przypadku.

Podsumowując całe zadanie możemy powiedzieć, że gdy liczba parametrów p jest bliska liczbie obserwacji n , to ciężko wybrać najlepsze kryterium. Możemy jedynie się posłużyć wnioskami z przypadku dla $n = 1000$. Natomiast, gdy liczba obserwacji n znacznie przewyższa liczbę parametrów p , to ewidentnie widzieliśmy różnicę. Kryteria RIC, mBIC i mBIC2 okazywały się dużo lepsze w tym przypadku, gdzie najlepszym z nich było kryterium mBIC.