

# Raport 1

Aleksander Milach

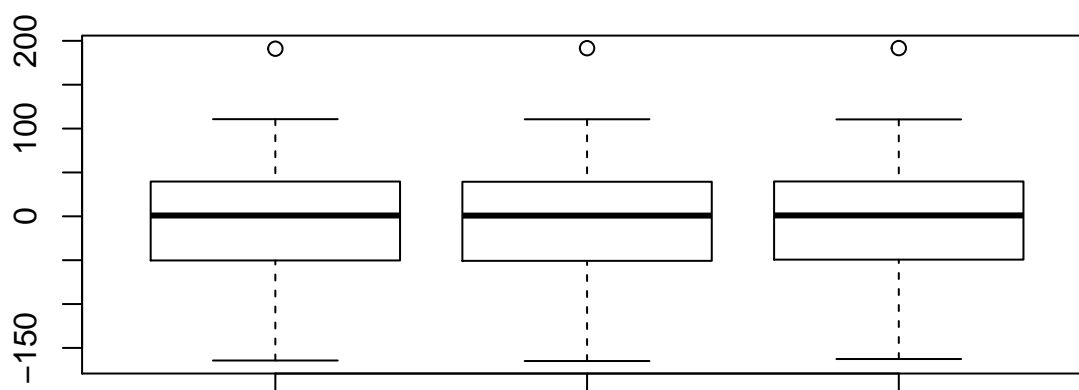
25 kwietnia 2019

## Zadanie 1

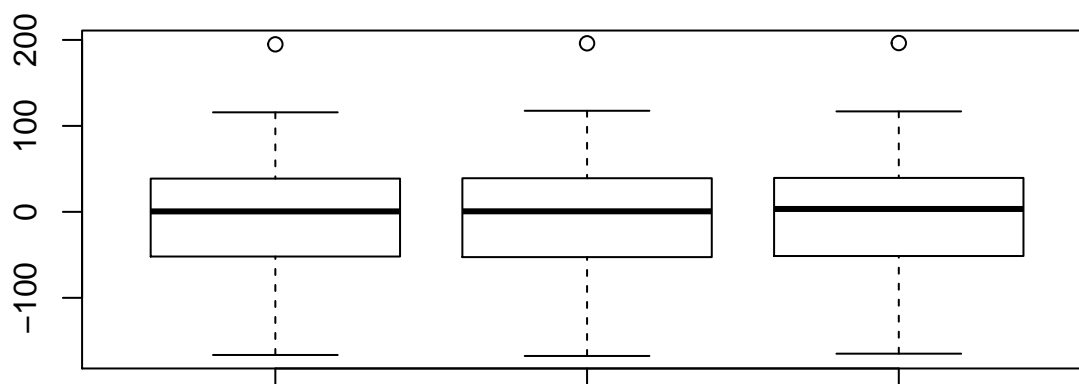
	5	10	20	100	500	950
	zmiennych	zmiennych	zmiennych	zmiennych	zmiennych	zmiennych
RSS	984.5300	979.2134	974.8128	918.1789	531.4624	58.01536
PE	1081.9161	1086.4298	1089.4699	1148.3036	1965.4556	16976.87345
estPE1(AICzn)	994.5300	999.2134	1014.8128	1118.1789	1531.4624	1958.01536
estPE2	994.4348	999.0155	1014.6417	1122.4456	1596.5174	2307.59051
estPE3	995.0509	999.6243	1015.2666	1133.6922	2124.8758	24828.76173
AICnz	994.4248	998.9955	1014.6011	1122.2186	1594.3873	2262.59901

W obu przypadkach AIC, dla znanego i nieznanego sigma, AIC wybiera model z 5 zmiennymi.

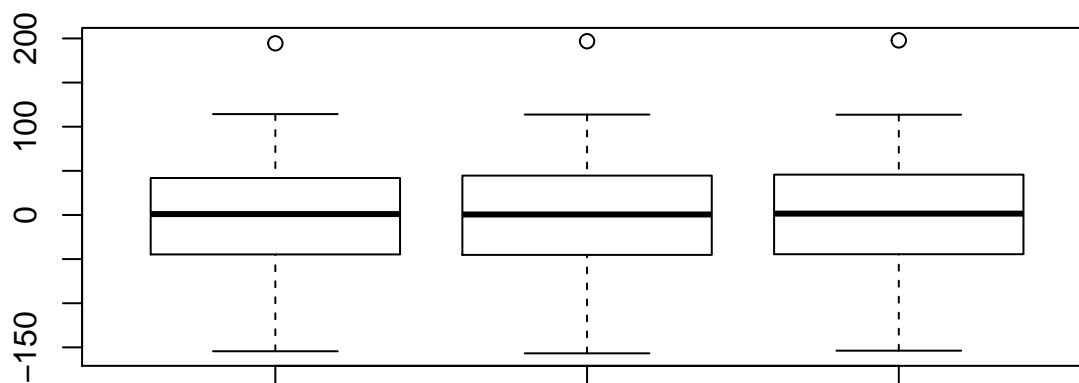
### 5 pierwszych zmiennych Estymatory-PE



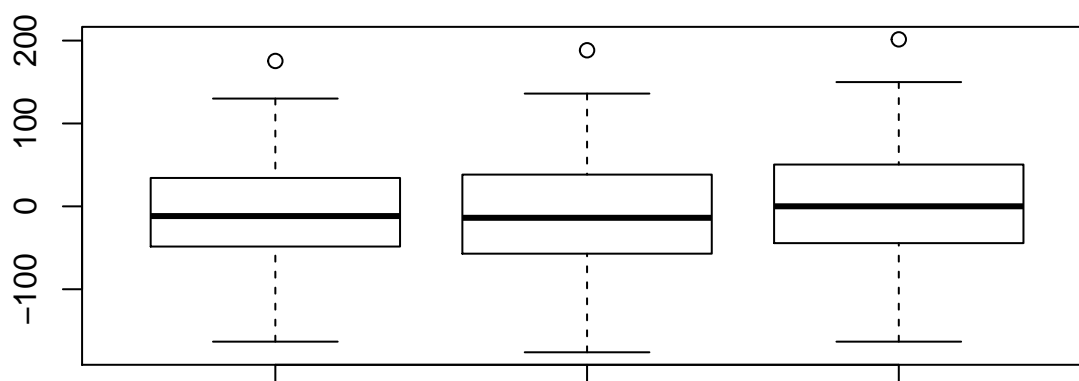
### 10 pierwszych zmiennych Estymatory-PE



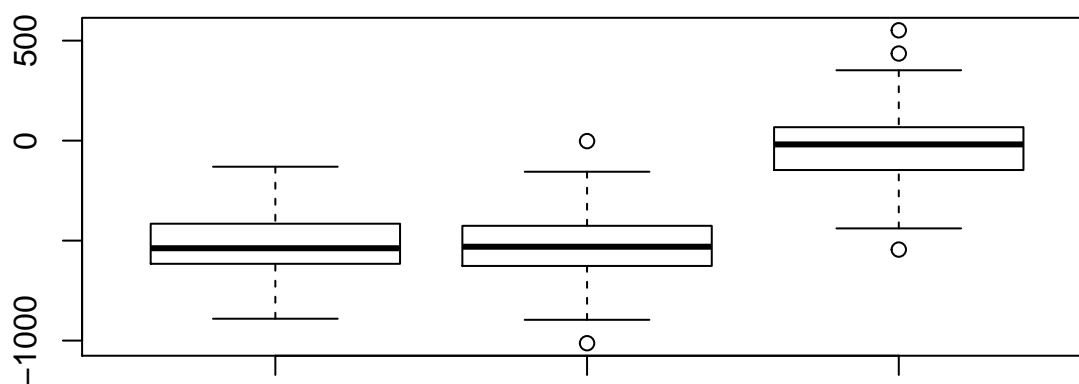
**20 pierwszych zmiennych Estymatory-PE**



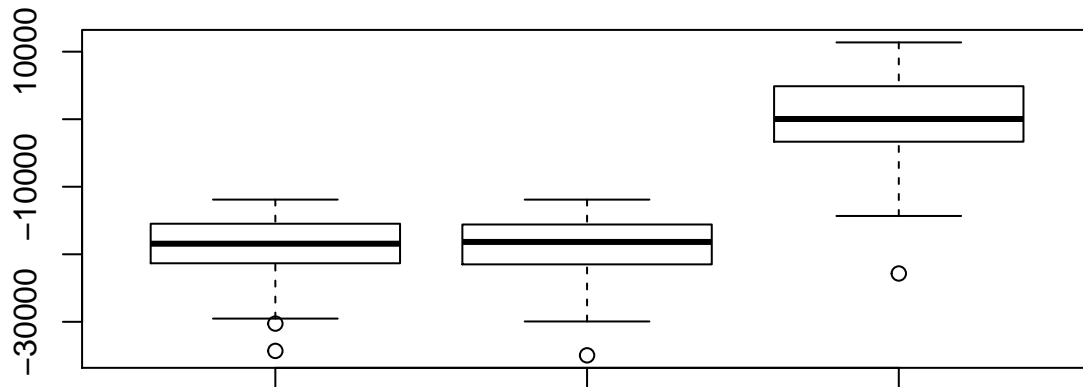
**100 pierwszych zmiennych Estymatory-PE**



**500 pierwszych zmiennych Estymatory-PE**



## 950 pierwszych zmiennych Estymatory-PE



Nawet gdy bierzemy 100 pierwszych zmiennych wszystkie estymatory błędów predykcji mają średnią zero i rozsądną wariancję. W przypadku gdy  $n=500$  lub  $n=950$  estymatory związane z RSS zaczynają zaniżać błąd predykcji, jest to przykład na przepasowanie modelu do danych; gdy  $p$  zbliża się do  $n$ , model dopasowuje się do losowego błędu.

### Zadanie 2

#### Podpunkt a

	Ta	Fa	SEa	Tb	Fb	SEb	Tm	Fm	SEm	Tm2	Fm2	SEm2
20 zmiennych	1	0.2857143	0.0082718	1	0.0000000	0.0065567	1	0	0.0065567	1	0	0.0065567
100 zmiennych	1	0.7619048	0.0240427	1	0.0000000	0.0065567	1	0	0.0065567	1	0	0.0065567
500 zmiennych	1	0.9285714	0.0589258	1	0.4444444	0.0091287	1	0	0.0065567	1	0	0.0065567
950 zmiennych	1	0.9285714	0.0589258	1	0.6428571	0.0111702	1	0	0.0065567	1	0	0.0065567

#### Podpunkt b

	20 zmiennych	100 zmiennych	500 zmiennych	950 zmiennych
Pow. AIC	1.0000000	1.0000000	1.0000000	1.0000000
FDR AIC	0.2919625	0.7375462	0.9284921	0.9285714
MSE AIC	0.0085687	0.0200553	0.0659328	0.0659661
Pow. BIC	1.0000000	1.0000000	1.0000000	1.0000000
FDR BIC	0.0195238	0.1263690	0.4382295	0.6164542
MSE BIC	0.0066171	0.0071899	0.0103679	0.0147977
Pow. mBIC	1.0000000	1.0000000	1.0000000	1.0000000
FDR mBIC	0.0033333	0.0083333	0.0083333	0.0100000
MSE mBIC	0.0065707	0.0065759	0.0065911	0.0065939
Pow. mBIC2	1.0000000	1.0000000	1.0000000	1.0000000
FDR mBIC2	0.0373810	0.0411905	0.0457143	0.0457143
MSE mBIC2	0.0067537	0.0066456	0.0068802	0.0068436

Każde z kryteriów znalazło wszystkie zmienne znaczące, jednak w szczególności AIC i BIC dla  $n=500$  i  $n=950$  mamy bardzo dużo fałszywych odkryć. Znacznie mniej fałszywych odkryć wybierało mBIC i mBIC2, dla każdego z czterech modeli.

### Zadanie 3

```
## [1] "6" "18" "28" "43"
```

```
## [1] "5" "6" "13" "18" "28" "29" "32" "43"
```

```
## [1] "5" "6" "13" "18" "28" "29" "32" "43"
```

### Zadanie 4

#### Podpunkt a

	Tm	Fm	Tm2	Fm2	Trm	Frm	Trm2	Frm2
Shifted exp.	29	0	30	1	29	0	30	1
Cauchy	0	0	0	0	5	0	9	2

W przypadku rozkładu wykładniczego, każde z kryteriów zachowuje się podobnie, znajduje znaczną większość zmiennych istotnych i mało lub żadnych fałszywych odkryć.

Dla rozkładu Cauchy'ego kryteria mBIC i mBIC2 nie znajdują żadnych zmiennych istotnych, błąd o rozkładzie Cauchy'ego jest 'za trudny' dla zwykłych kryteriów. Kiedy zastosujemy podejście rangowe kryteria znajdują część (około 1/4) zmiennych istotnych, bez fałszywych odkryć.

#### Podpunkt b

	LM	Huber	Bsquare
Shifted exp.	2.935449	2.89098	2.900609
Cauchy	2.501440	2.48765	2.490061

Powyżej wymieniałem wartości MSE dla każdego z eksperymentów.

#### Podpunkt c

	Pow. mBIC	FDR mBIC	Pow. mBIC2	FDR mBIC2	Pow. rBIC	FDR rBIC	Pow. rBIC2	FDR rBIC2
Exp.	0.9	0	0.9333333	0.0010526	0.9333333	0	0.9333333	0.0021053
Cauchy	0.0	0	0.0000000	0.0000000	0.2000000	0	0.3000000	0.0010526

Dla błędu o rozkładzie wykładniczym mamy bardzo pożądane własności, moc powyżej 0.9 i sporadyczne fałszywe odkrycia. Dla błędu o rozkładzie Cauchy'ego kryteria nierangowe nie znajdują żadnych zmiennych, zaś rangowe znajdują niektóre zmienne istotne, ale wciąż bardzo mały ich ułamek.