# Lista 3

*Aneta Przydróżna*

**Porównanie działania regresji logistycznej dla różnych warunków początkowych.**

- Prawdziwe odkrycia
- Fałszywe odkrycia
- FDR
- $\beta(\hat{\beta}) = \frac{1}{p}\sum_{i=p}^{p}(\hat{\beta}_i - \beta_i)$
- $SE(\hat{\beta}) = \| \hat{\beta} - \beta \|^2$
- $SE(\hat{p}) = \| \hat{p} - p \|^2$

Poniżej znajdują się uśrednione wartości tych statystyk z 20 badań dla standardowej regresji logistycznej, zgodnych z kryteriami BIC, mBIC, mBIC2 modeli oraz modelu regresji logistycznej z użyciem SLOPE i LASSO z walidacją krzyżową.

## $X_{1000\times250} \sim N(0,1)$ **oraz** $\beta = (c,c,c,c,c,0,..,0)^T$**, gdzie** $c = \frac{9}{\sqrt{n}}$

```
##          Bias SE(B) SE(p) True False   FDR
## glm    0.0029  2.99  6.96 4.95  23.4 0.819
## BIC   -0.0043  0.25  6.62 3.95   0.5 0.094
## mBIC  -0.0024  0.12  6.79 2.75   0.0 0.000
## mBIC2 -0.0029  0.15  4.94 3.25   0.0 0.000
## LASSO -0.0053  0.32 15.72 4.93  15.8 0.701
## SLOPE  0.0004  0.09  4.15 3.99   0.3 0.061
```

W klasycznej regresji logistycznej, błąd kwadratowy jest największy ze wszystkich przypadków, również fałszywych odkryć jest bardzo dużo (średnio 23), co wpływa na wysoki wskaźnik FDR, reszta statystyk nie różni się znacząco od pozostałych metod. Lasso wyróżnia się największym obciążeniem (choć stosunkowo, wciąż niedużym), dużą ilością fałszywych odkryć i dużym błędem predykcji prawdopodobieństw. Obie metody nie wypadły najlepiej, natomiast odkryły największą ilość prawdziwych współczynników, z czym nie poradził sobie za dobrze mBIC, przez surową karę, która za to nie dopuszcza fałszywych odkryć, mBIC2 poradził sobie trochę lepiej z odkryciami, a z resztą statystyk podobnie. BIC nie dopuszczał za dużo fałszywych oraaz dosyć dobrze wykrywał prawdziwe, za to ma większe obciążenie i błąd kwadratowy. Slope łapał mało fałszywych odkryć, miał małe obciążenie i błędy, a jego wynik w wyszukiwaniu prawdziwych odkryć jest jednym z lepszych.

## $X_{1000\times250} \sim N(0,1)$ **oraz** $\beta = (c,..,c,0,..,0)^T$**, gdzie** $c = \frac{9}{\sqrt{n}}$**, powtórzone 50 razy**

```
##          Bias SE(B) SE(p) True False  FDR
## glm    0.0483 11.68  54.6 45.6  27.6 0.37
## BIC   -0.0310  1.85  51.9 22.0   1.5 0.06
## mBIC  -0.0058  2.32  83.5  6.1   0.0 0.00
## mBIC2 -0.0130  2.74  73.7 11.9   0.0 0.01
## LASSO -0.0043  3.32  65.9 49.0  75.7 0.52
```

```
## SLOPE  0.0001  2.09  47.8 26.1   1.3 0.06
```

Znowu największe SE przypada klasycznej metodzie, która bardzo dobrze poradziła sobie ze znalezieniem prawdziwych odkryć, ze stratą podobną do przypadku z mniejszą liczbą istotnych zmiennych, a więc lepiej. SE(p) znacząco wzrosło, jak w każdej metodzie. Lasso również świetnie sobie poradził z TD, niestedy wielkim kosztem FD. mBIC i mBIC2 prawie bezbłędnie odrzucały FD, przy bardzo złych wynikach w odnajdywaniu TP. Slope z najmniejszymi błędami i obciążeniem, radziło sobie przexiętnie z TD, przy bardzo małej stracie.

## $X_{1000\times250} \sim N(0,\Sigma)$, gdzie $\Sigma$ to macierz z 1 na przekątnej i 0.3 w pozostałych miejscach oraz $\beta = (c,c,c,c,c,0,..,0)^T$, gdzie $c = \frac{9}{\sqrt{n}}$

```
##          Bias SE(B) SE(p) True False  FDR
## glm    0.0032  5.43   7.2  4.5  24.3 0.82
## BIC    0.0000  0.16   4.8  3.6   1.9 0.31
## mBIC   0.0001  0.10   3.0  3.0   0.0 0.00
## mBIC2  0.0010  0.08   2.4  4.0   0.1 0.02
## LASSO -0.0056  0.33  17.9  4.9  21.5 0.78
## SLOPE  0.0022  0.17   6.2  4.0  29.8 0.85
```

We wszystkich przypadkach obciążenie jest dość małe, SE(B) znowu wyróżnia się tylko w klasycznym glm. SE(p) jest największe dla Lasso, które znowu najlepiej szuka odkryć, choć z dużym nadmiarem tych fałszywych. Slope wcześniej radził sobie dobrze , ponieważ nie dopuszczał fałszywych odkryć, w przypadku gdy macierz X jest skorelowana, znajduje ich bardzo dużo, z podobnym efektem dla prawdziwych odkryć. Dla kryteriów wskaźnik FDR jest bardzo mały, natomiast ilość prawdziwych odkryć przeciętna.

## $X_{1000\times250} \sim N(0,\Sigma)$, gdzie $\Sigma$ to macierz z 1 na przekątnej i 0.3 w pozostałych miejscach oraz $\beta = (c,..,c,0,..,0)^T$, gdzie $c = \frac{9}{\sqrt{n}}$, powtórzone 50 razy

```
##         Bias    SE(B) SE(p) True False  FDR
## glm   -0.052   432.10 50.60  0.0   0.0 0.72
## BIC    0.008     2.00 20.50 29.3   1.9 0.05
## mBIC  -0.140     1.12 26.90 19.0   0.1 0.01
## mBIC2  0.004     1.68 20.20 22.5   1.1 0.03
## LASSO  0.027     4.04 11.83 37.5  34.0 0.54
## SLOPE -3.532 42338.20 53.20 47.9 129.5 0.00
```

Tym razem Slope nie poradziło sobie z błędami i obciążeniem, w pewnym momencie pracy nad danymi rosły do wielkich rozmiarów, a przy wysokiej skuteczności szukania TD, znalazł bardzo dużo FD. Można zauważyć, że slope działa dobrze, gdy elementy macierzy X są niezależne. Klasyczna metowa w ogóle sobie nie poradziła. Wszystkie statystyki są zawyżone, a ilość odkryć to zero. Lasso wypadło poprawnie. z Kryteriów chyba najlepiej poradził sobie BIC, jeśli przyjęcie szumu nie jest dla nas aż tak dotkliwe.

**Tym razem badamy także zbadamy własności regresji Firtha, w wersji eksperymentu z macierzą 100 na 25.**

$X_{100\times25} \sim N(0,1)$ **oraz** $\beta = (c,c,c,c,c,0,..,0)^T$**, gdzie** $c = \frac{9}{\sqrt{n}}$

```
##          Bias   SE(B) SE(p) True False  FDR
## Firth  0.0020    3.70   3.6 3.80  0.75 0.10
## OLS    1.2290 1293.00   6.7 4.00  2.10 0.39
## LASSO -0.1480    3.42   8.3 4.95  6.10 0.60
## SLOPE -0.0030    1.40   2.3 3.70  0.30 0.07
## mBIC2 -0.0021    2.80   2.9 3.95  0.60 0.15
```

Wyniki metody najmniejszych kwadratów najbardziej się wyróżniają, pokazując jej wady. Błąd kwadratowy wyszedł ogromny, bias również jest o wiele większy od pozostałych. Znajdowanie prawdziwych odkryć idzie nieźle, ale niesie ze sobą wskaźnik FDR. Lasso wypadło słabo we wszystkich wynikach, poza TD, gdzie uzyskało najwyższy, prawie maksymalny wskaźnik. Z pozostałej trójki ciężko wybrać najlepszą metodę, ponieważ każda wypada w czymś najlepiej. Należałoby to rozstrzygnąć biorąc pod uwagę najbardziej pożądane kwestie.

$X_{100\times25} \sim N(0,\Sigma)$**, gdzie** $\Sigma$ **to macierz z 1 na przekątnej i 0.3 w pozostałych miejscach oraz** $\beta = (c,c,c,c,c,0,..,0)^T$**, gdzie** $c = \frac{9}{\sqrt{n}}$

```
##        Bias  SE(B) SE(p) True False  FDR
## Firth  0.01    4.3   3.6  2.7   0.8 0.20
## OLS   32.00 4343.0   6.9  2.0   1.5 0.56
## LASSO -0.16    3.6   7.6  4.9   5.1 0.50
## SLOPE  0.09    1.1   2.3  4.5   4.1 0.41
## mBIC2  0.03    3.7   2.8  3.8   0.6 0.16
```

Po wprowadzeniu niezależności wektorów macierzy X, metoda OLS działa jeszcze gorzej, ma większe błędy, dwa razy gorszy wskaźnik TD oraz wyższą wartość FDR. Regresja Firtha ogólnie działa gorzej. Lasso i Slope dobrze odnajdują TP, niestety z dużym zapasem FN. mBIC2 poradził sobie w tym wypadku podobnie jak wyżej. Tym razem również bardzi ciężko rozstrzygnąć która metoda wypadła najlepiej. Podsumowując wszyskie wyniki, można stwierdzić, przetwarzając dane przez kryterium mBIC, kolejno przez regresję logistyczną otrzymujemy, w każdym przypadku, podobne wskaźniki. Zawsze były to niskie liczby w obciążeniu i SE oraz bardzo niski wskaźnik FDR, jedynym minusem metody jest średnia moc.