# Selekcja cech

Wojciech Kosztyła, Wojciech Węgrzynek

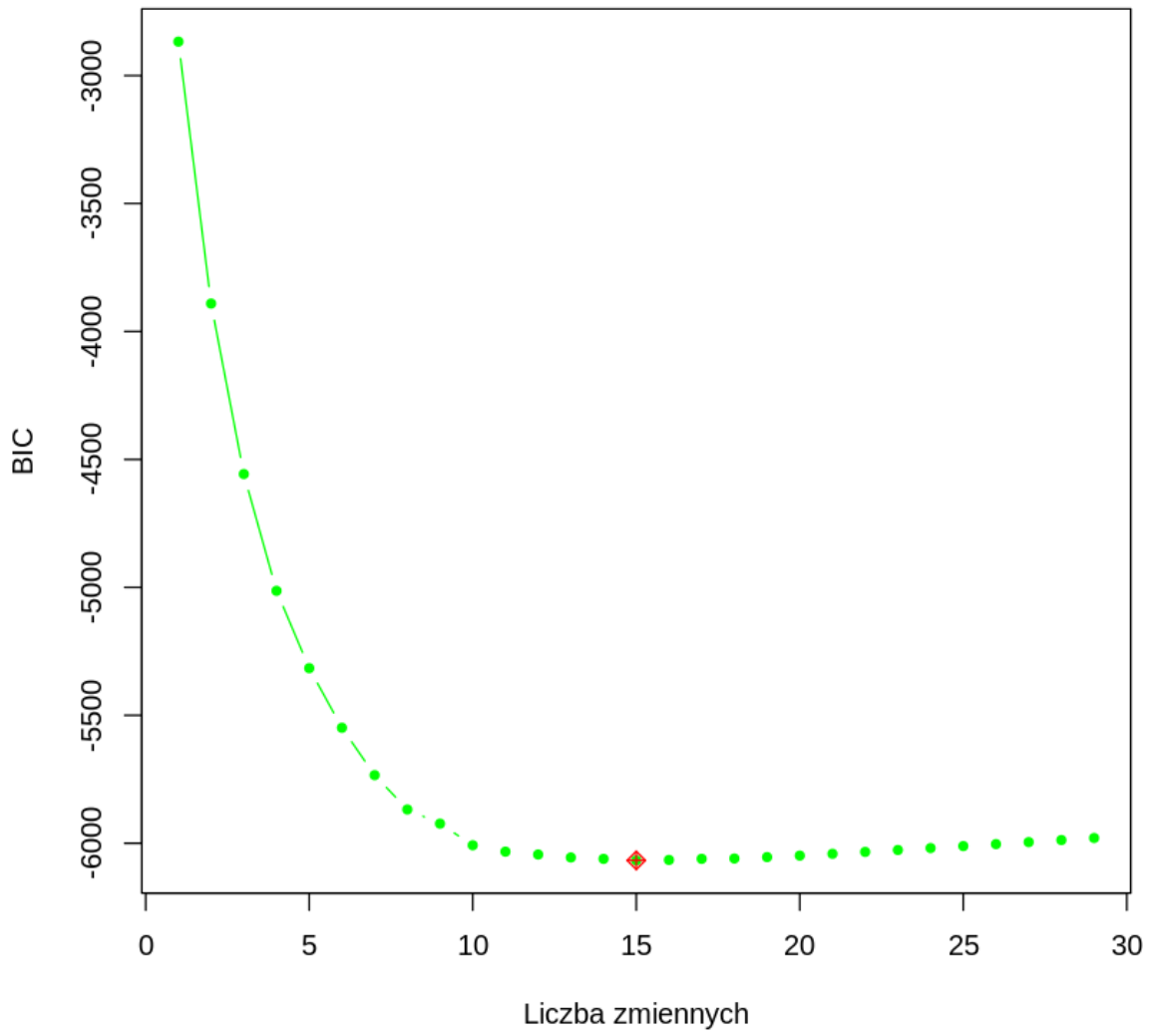
## Regresja

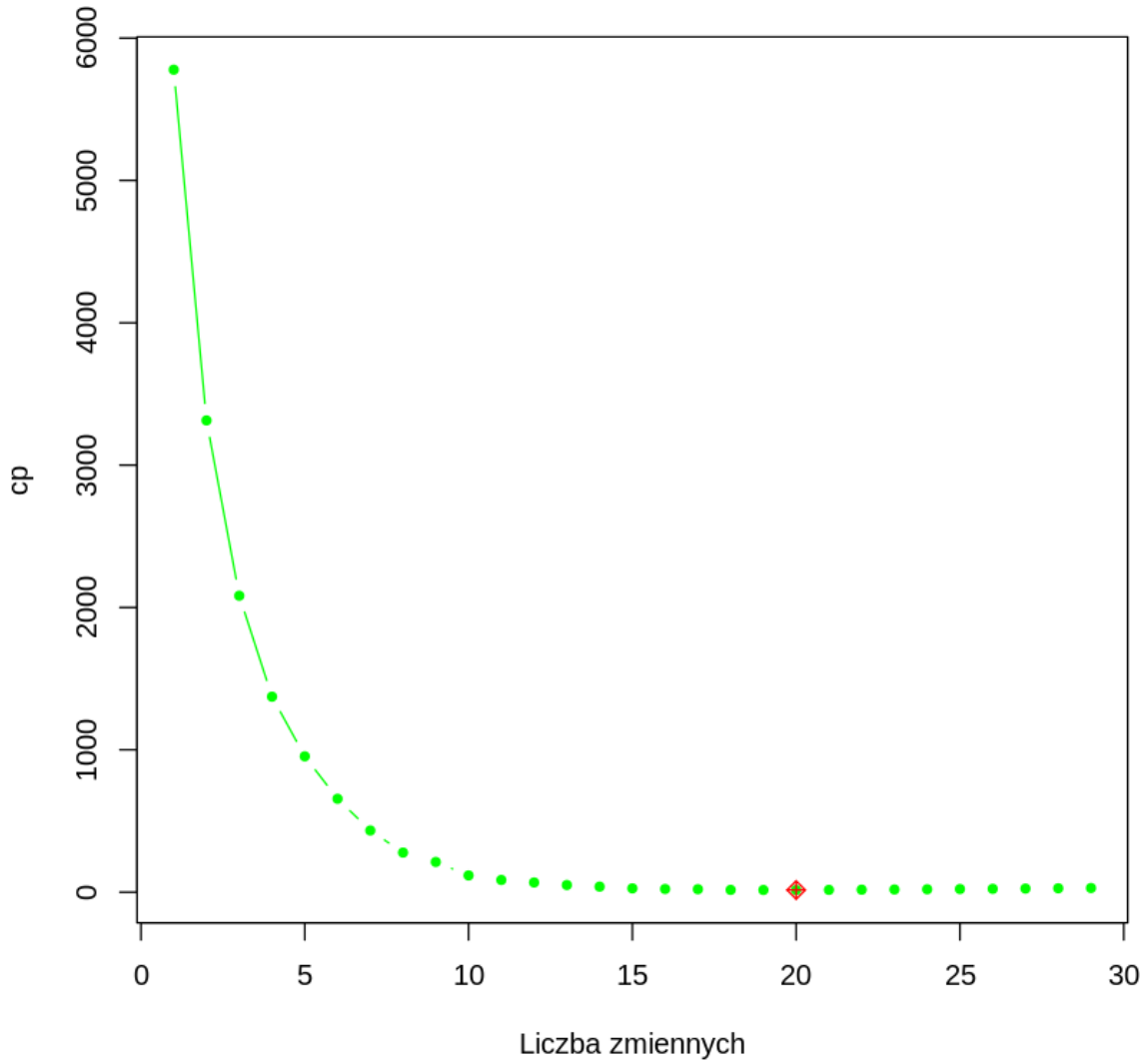
Dla przypomnienia, w ramach regresji przewidujemy medianę dochodu gospodarstw domowych na podstawie danych o hrabstwach z censusu amerykańskiego 2015.

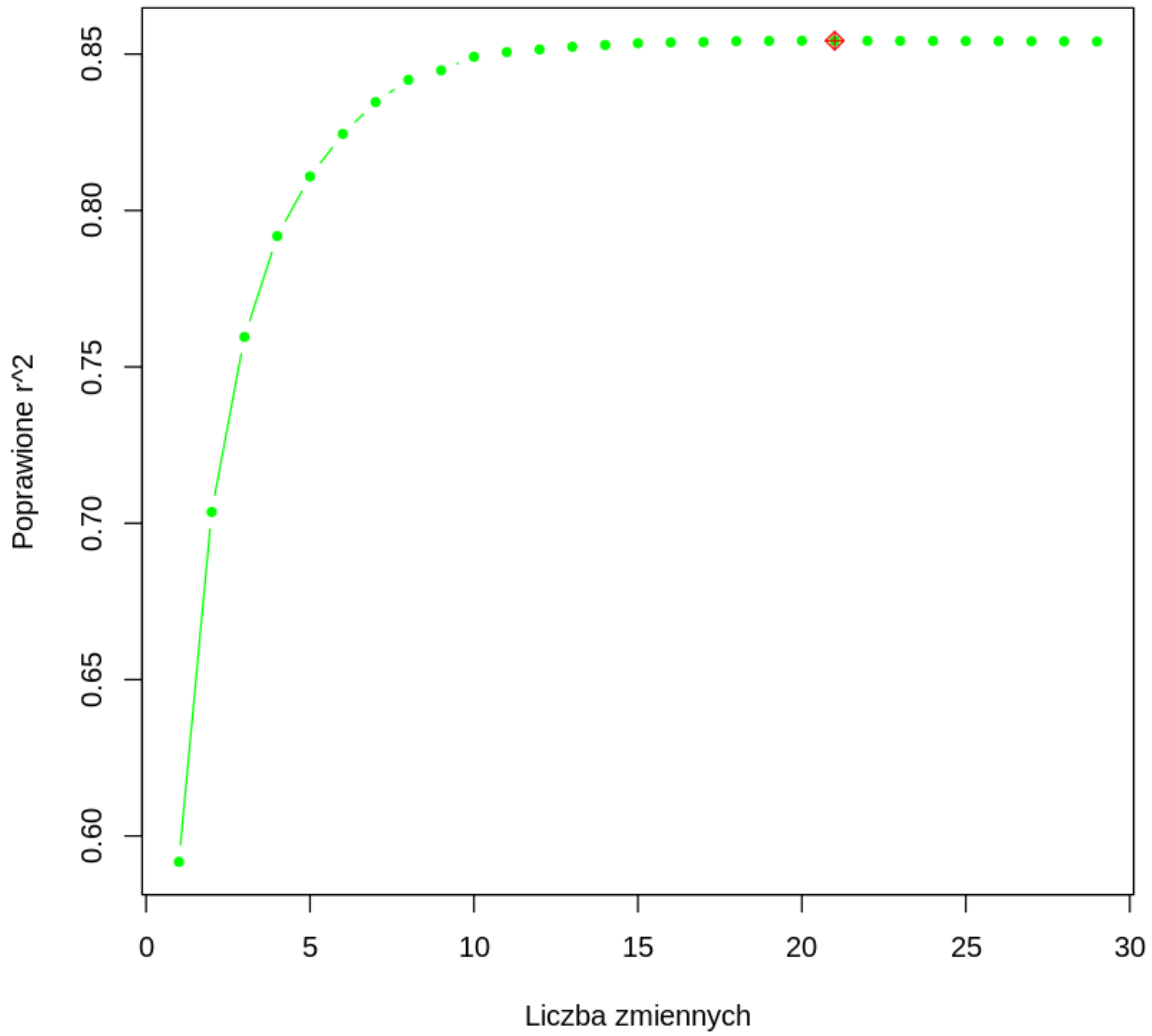
### Regsubsets

Na początku przeprowadziliśmy selekcję cech funkcją regsubsets, przeglądając wszystkie 29 cech. Co ciekawe, czas wykonania był niezauważalny dla człowieka (zaskakująco mały).

Wyboru liczby cech dokonaliśmy za pomocą miar: `bic`, `cp` oraz poprawionego `r^2`. Poniżej zamieszczone są wykresy tych miar (na czerwono zaznaczone są odpowiednie ekstrema).




cp
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25
30
Liczba zmiennych




Ponieważ ekstrema tychże miar wskazywały wciąż na dużą liczbę cech, zdecydowaliśmy się wybrać mniejszą liczbę cech. Poświęciliśmy w ten sposób trochę dokładności modelu na rzecz jego czytelności.

Dobrym kandydatem na liczbę cech wydaje się być 10, jedynie cp odstaje dla niego znacząco od ekstremum.

| **Liczba cech** | `BIC` | `cp` | `adjr2` |
|---|---|---|---|
| 10 | -6007.64942041253 | 117.337011847296 | 0.849137068465889 |
| ekstremalna | -6066.69853851265 | 14.8681687559433 | 0.854261601115093 |

Za najbardziej istotne predyktory można więc uznać te, które pojawiły się w optymalnych modelach dla 10 (lub mniej) cech. Takie predyktory to:

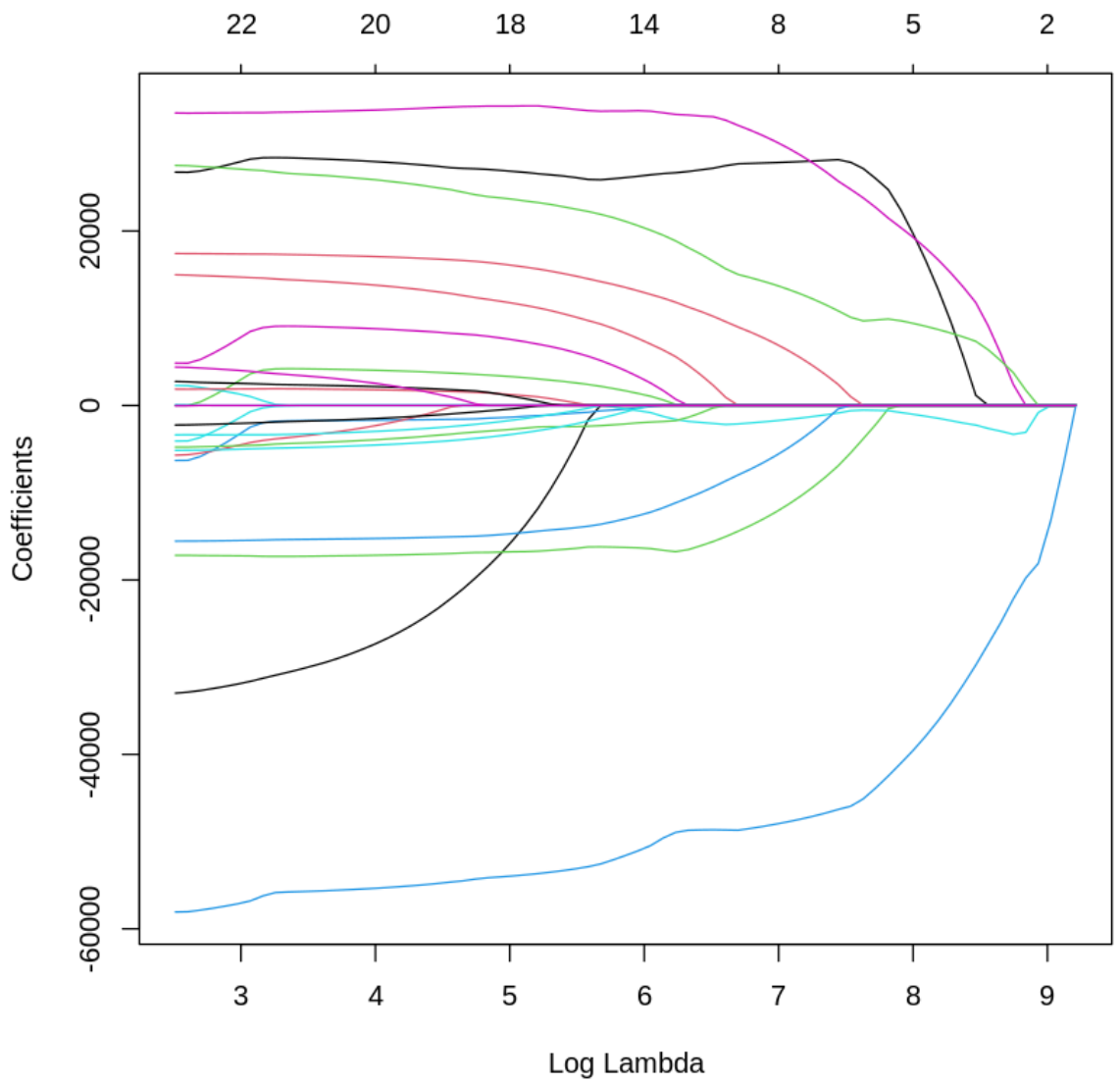
- `Poverty` - procent ludności poniżej granicy ubóstwa. Używany dla każdej liczby predyktorów.

- `Professional` - procent ludności zatrudnionej w zarządzaniu, biznesie, nauce i sztuce. Używany dla każdej liczby predyktorów (oprócz 1).
- `Citizen` - procent ludności posiadający obywatelstwo Stanów Zjednoczonych. Używany dla każdej liczby predyktorów (oprócz 1 i 2).
- `SelfEmployed` - procent ludności na samozatrudnieniu.
- `MeanCommute` - średni czas przejazdów.
- `Asian` - procent ludności etniczności azjatyckiej.
- `Employed` - procent ludności pracującej.
- `Men`/`Women` - procent ludności danej płci (cechy wykryte jako liniowo zależne, model regsubsets czasem wybiera jedną, czasem drugą).
- `Native` - procent ludności rdzennych amerykanów.
- `Hispanic` - procent ludności etniczności latynoamerykańskiej.

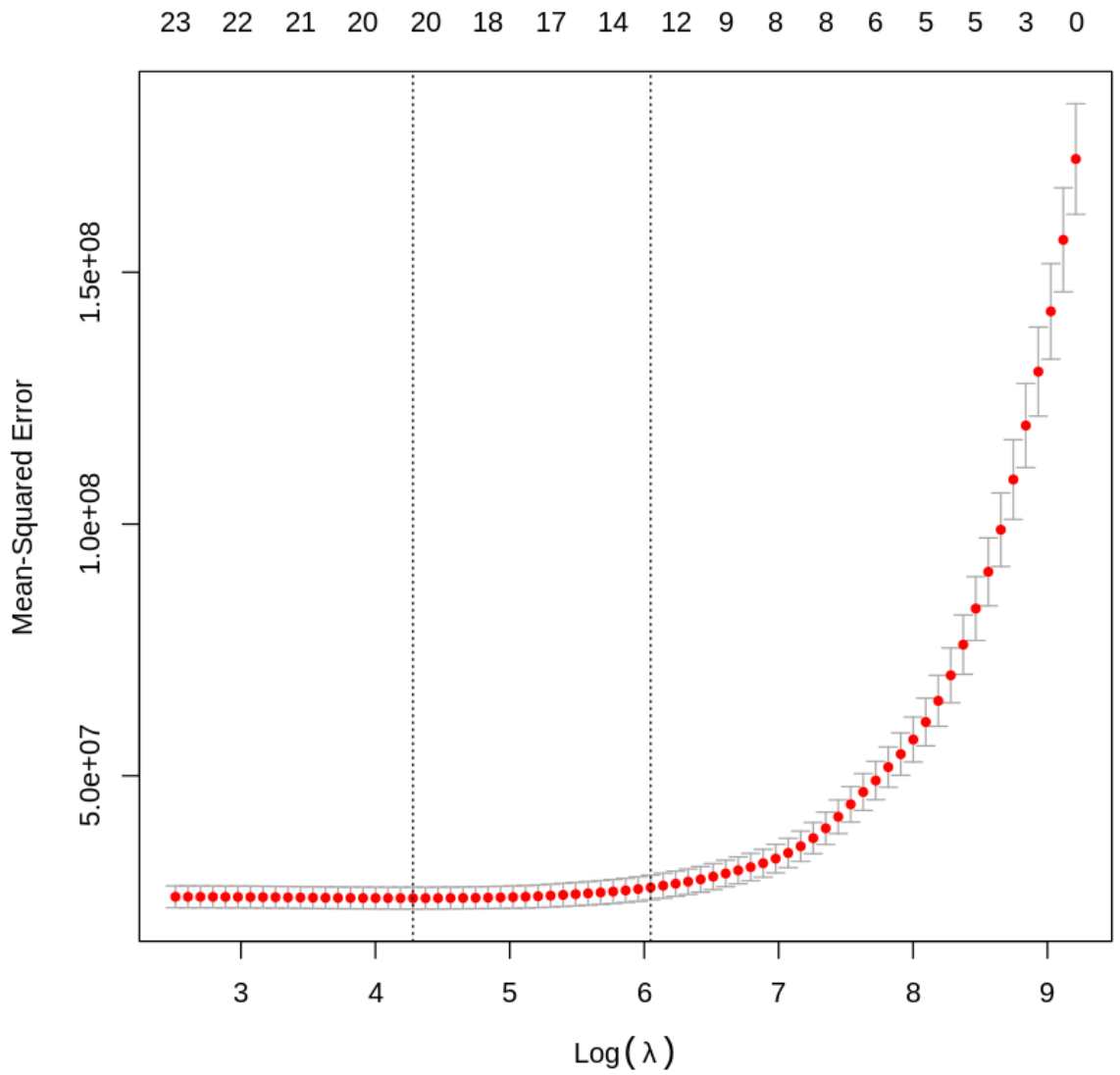
W metodach skokowych używane cechy były bardzo podobne. W metodzie `forward` zostały użyte dokładnie te same cechy (przynajmniej dla naszej liczby cech). W metodzie `backward` zaniechane zostały predyktory `Native` i `Hispanic` na rzecz `White` (procent populacji białej), oraz `Black` (procent populacji czarnej).

## Regresja lasso

Następnie wykonaliśmy regresję lasso w celu wyznaczenia ważnych cech. Wykres poniżej przedstawia zależność współczynników (na znormalizowanych danych) od lambdy.




Z kolei ten przedstawia zależność błędu na zbiorze testowym od lambdy, z zaznaczonym minimum. Można w szczególności zauważyć, że nasz model przeucza jedynie nieznacznie - model bez regularyzacji popełnia niewiele większy błąd, a niewielka stała lambda wystarczy aby ten błąd zminimalizować.




Jeżeli chodzi o selekcję cech za pomocą optymalnego lambda, odrzuciliśmy jedynie 9 z 30 cech. Cechy **odrzucone** przez takie lambda to:

- `Black` - procent ludności czarnej.
- `ChildPoverty` - procent dzieci poniżej granicy ubóstwa.
- `Transit` - procent ludności jeżdżącej (do pracy) środkami komunikacji publicznej.
- `Construction` - procent ludności pracującej w budownictwie.
- `Drive` - procent ludności jeżdżącej (do pracy) samochodem.
- `Walk` - procent ludności chodzącej (do pracy) pieszo.
- `OtherTransp` - procent ludności jeżdżacej (do pracy) innymi środkami transportu.
- `WorkAtHome` - procent ludności pracującej w domu.
- `PublicWork` - procent ludności pracującej w sektorze publicznym.

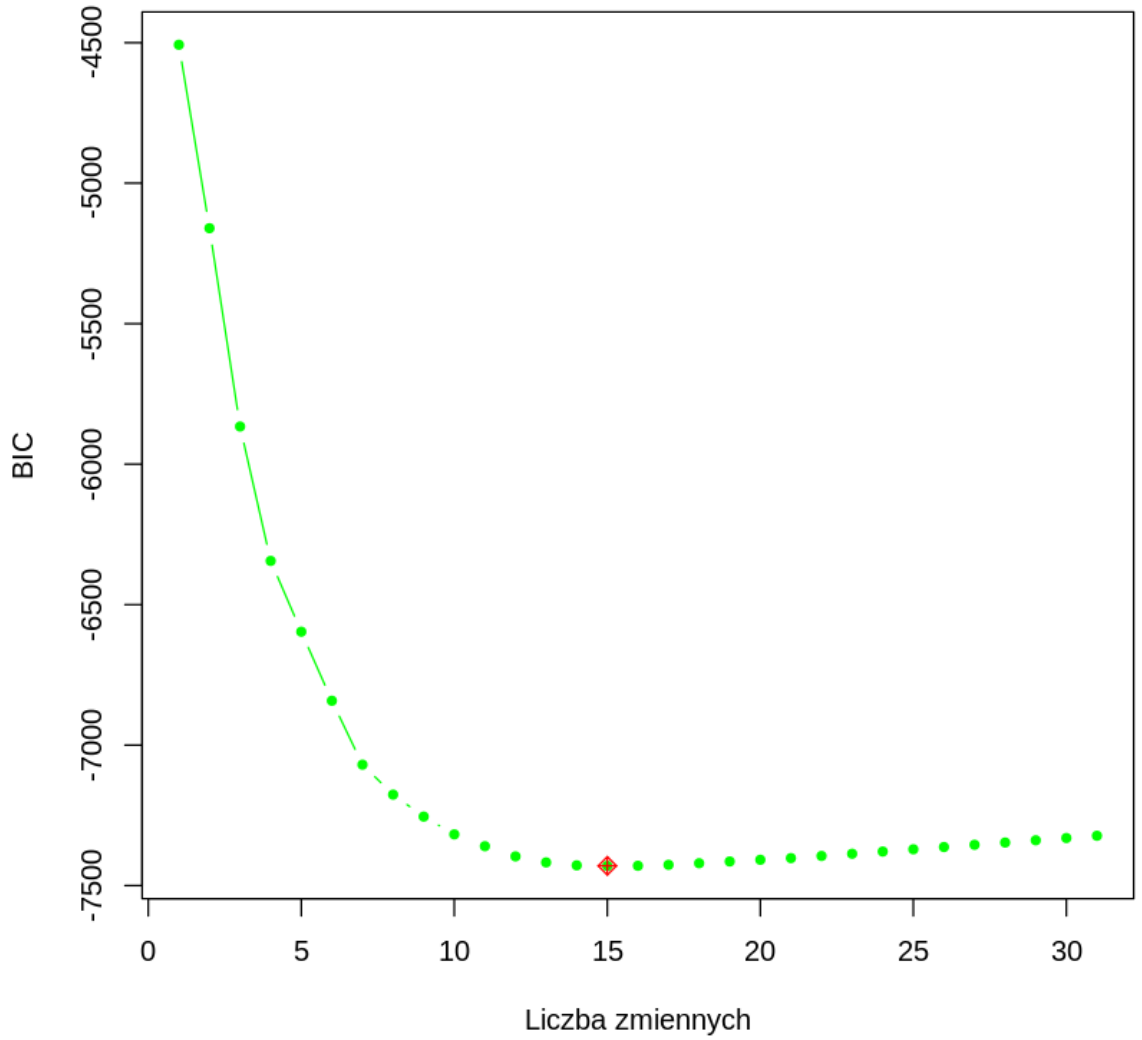
# Klasyfikacja

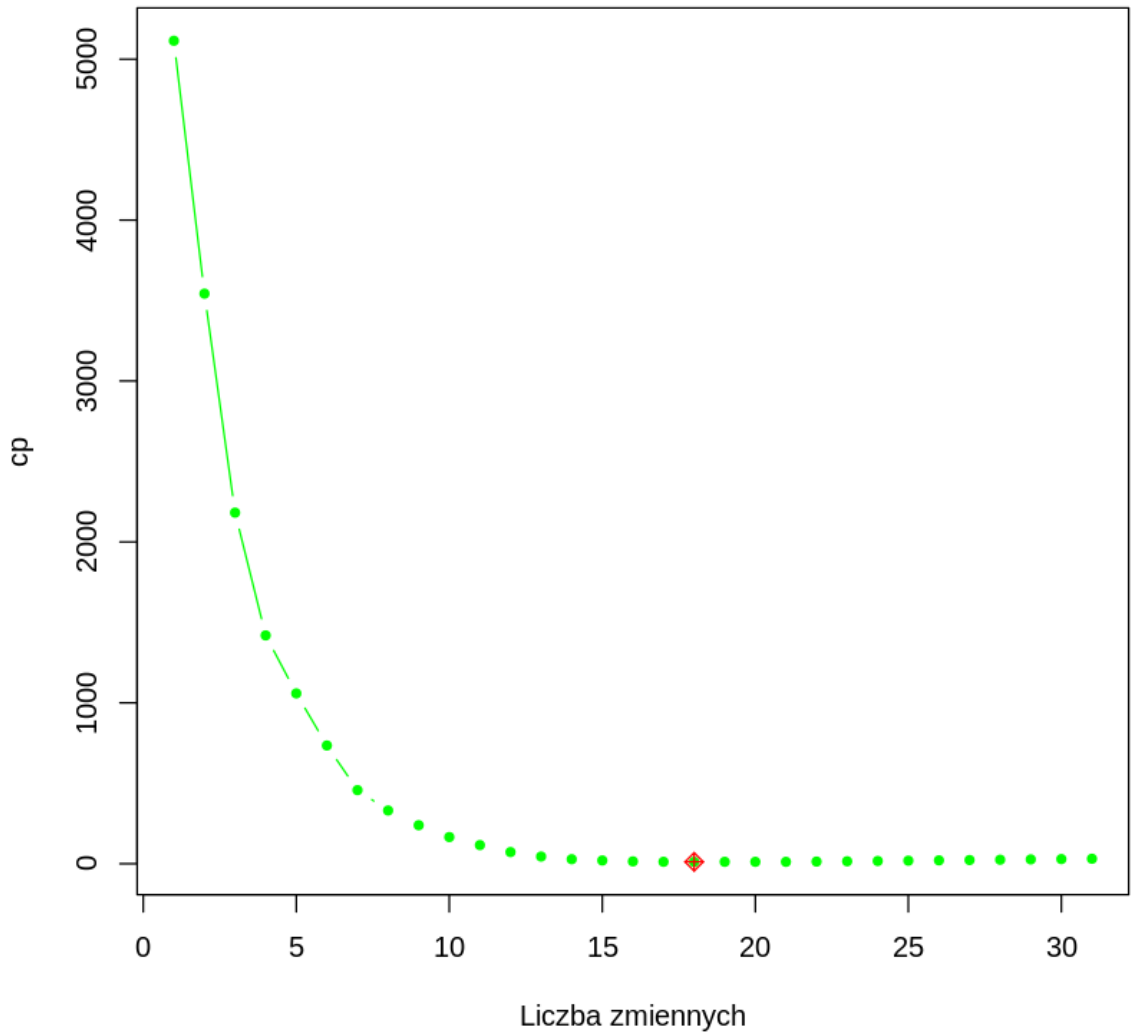
Proces dla klasyfikacji był analogiczny jak do regresji.
W ramach klasyfikacji przewidujemy zwycięzcę wyborów prezydenckich 2016 w danym hrabstwie.

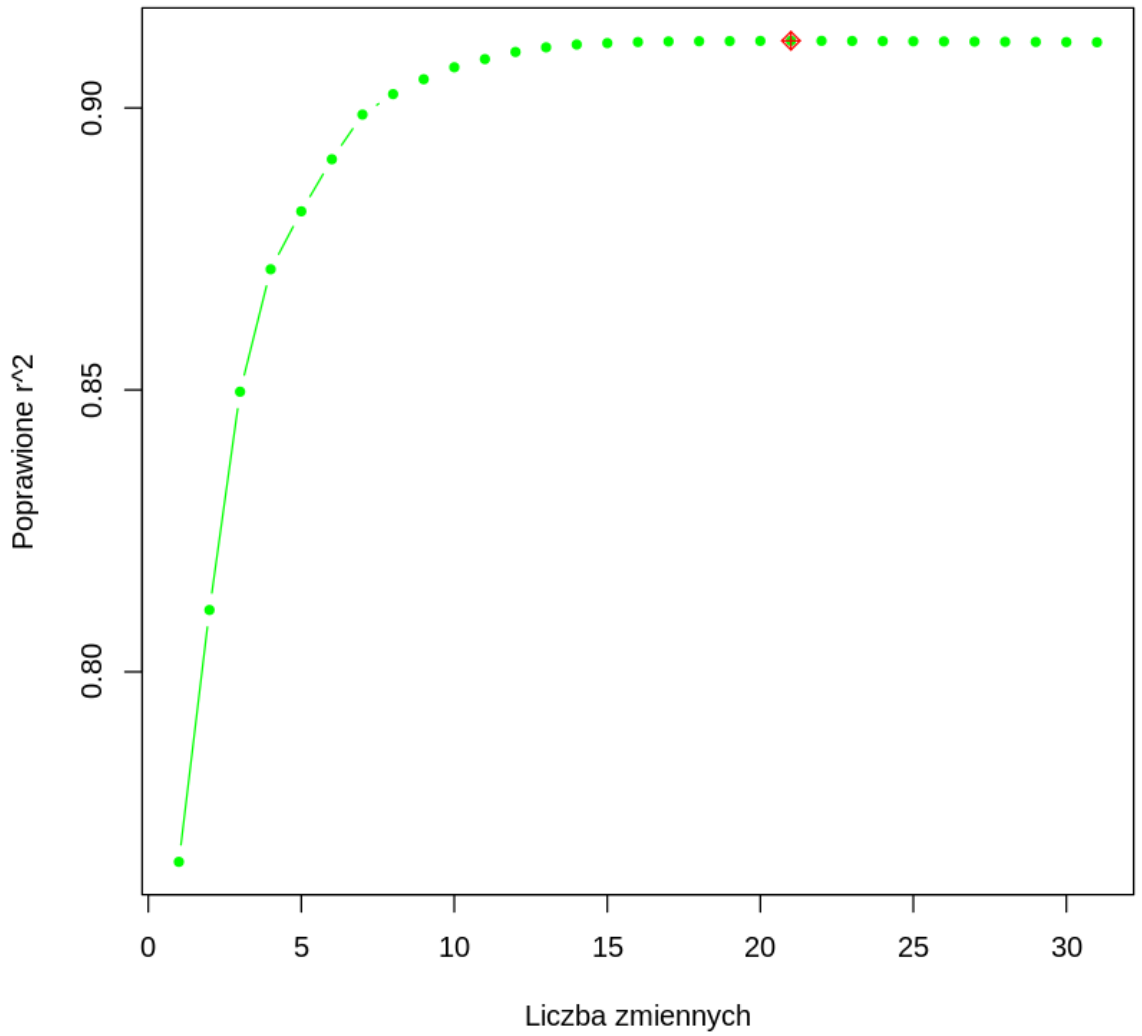
## Regsubsets

Selekcję cech funkcją regsubsets wykonaliśmy analogicznie, jak w przypadku regresji, tym razem jednak przeglądając 33 cechy.

Wyboru liczby cech ponownie dokonaliśmy za pomocą miar: `bic`, `cp` oraz poprawionego `r^2`. Poniżej zamieszczone są wykresy tych miar (na czerwono zaznaczone są odpowiednie ekstrema).




cp
0
1000
2000
3000
4000
5000
0
5
10
15
20
25
30
Liczba zmiennych




Ponieważ ekstrema tychże miar wskazywały wciąż na dużą liczbę cech, zdecydowaliśmy się wybrać mniejszą liczbę cech - 11.

| **Liczba cech** | `BIC` | `cp` | `adjr2` |
|---|---|---|---|
| 11 | -7359.71007779601 | 115.860506586702 | 0.908661718654433 |
| ekstremalna | -7429.70602258819 | 11.6952168437128 | 0.911913904000471 |

Za najbardziej istotne predyktory można więc uznać te, które pojawiły się w optymalnych modelach dla 11 (lub mniej) cech. Takie predyktory to:

- `Poverty` - procent ludności poniżej granicy ubóstwa. Używany dla każdej liczby predyktorów.
- `Professional` - procent ludności zatrudnionej w zarządzaniu, biznesie, nauce i sztuce. Używany dla każdej liczby predyktorów (oprócz 1).
- `Citizen` - procent ludności posiadający obywatelstwo Stanów Zjednoczonych.
- `SelfEmployed` - procent ludności na samozatrudnieniu.
- `MeanCommute` - średni czas przejazdów.

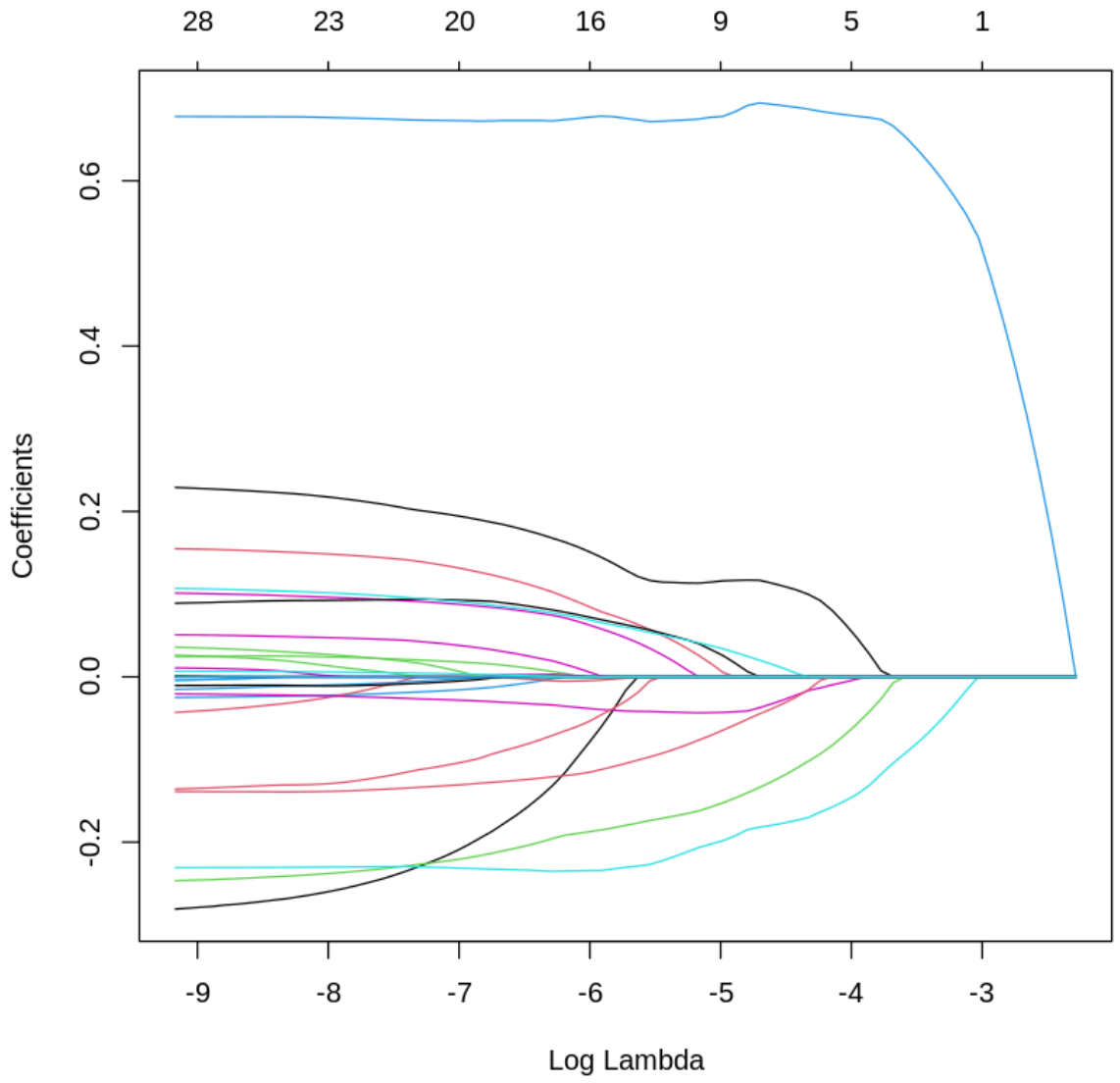
- `Asian` - procent ludności etniczności azjatyckiej.
- `Men`/`Women` - procent ludności danej płci (cechy wykryte jako liniowo zależne, model regsubsets czasem wybiera jedną, czasem drugą).
- `Native` - procent ludności rdzennych amerykanów.
- `Hispanic` - procent ludności etniczności latynoamerykańskiej.
- `IncomePerCapita` - średni dochód na osobę
- `Transit` - odsetek ludności podróżujących do pracy

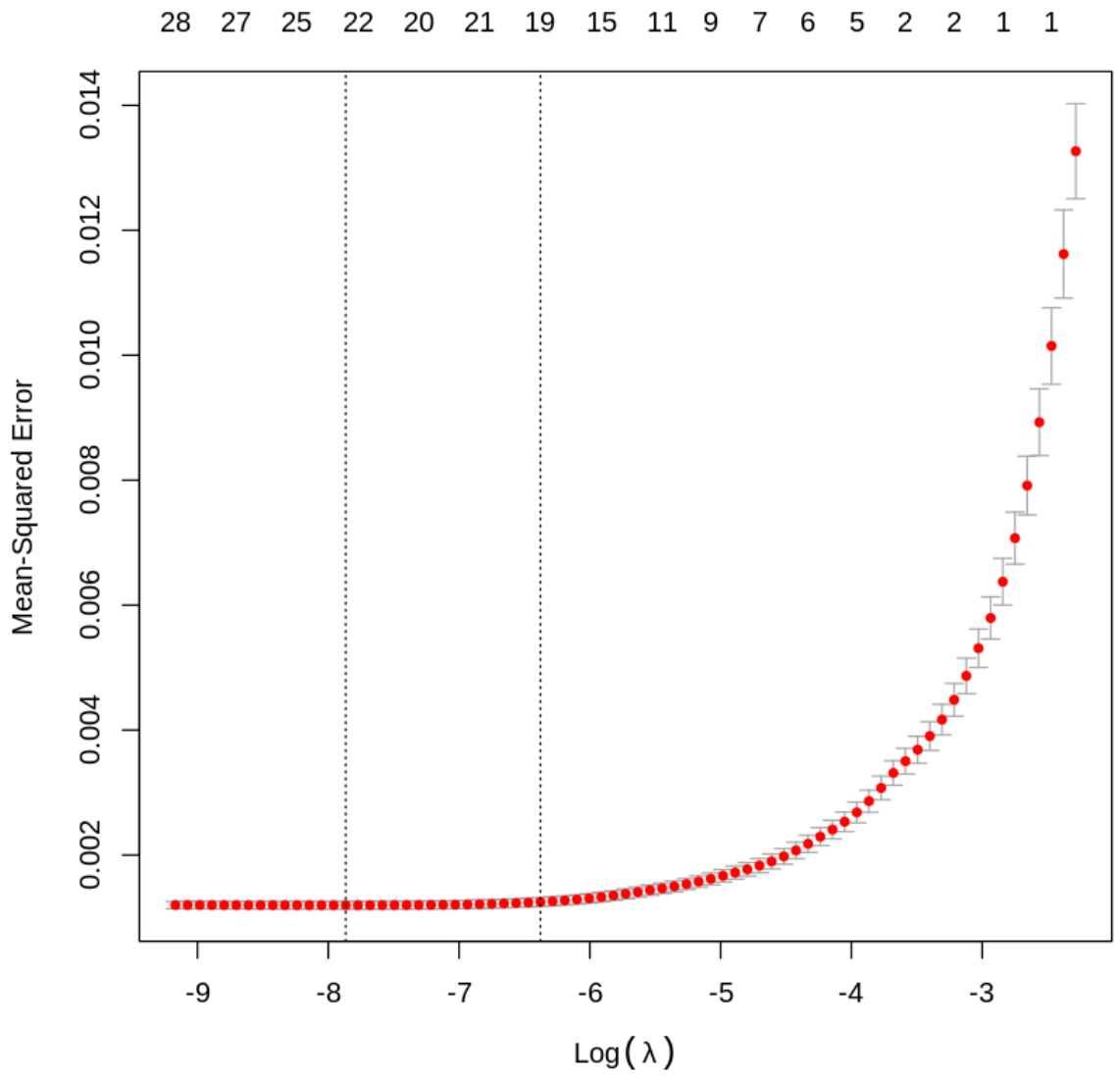
W porównaniu do predyktorów wskazanych przy regresji, różnią się brakiem `Employed` i pojawieniem się `IncomePerCapita` i `Transit`.

W metodach skokowych używane cechy były bardzo podobne. W metodzie `forward` zostały użyte dokładnie te same cechy, natomiast w metodzie `backward` zaniechany został predyktor `Transit` na rzecz `PrivateWork` (procent populacji pracującej z domu), co jak najbardziej ma sens.

## Regresja lasso

Następnie wykonaliśmy regresję lasso w sposób identyczny do poprzedniego, w celu wyznaczenia ważnych cech. Wykres poniżej przedstawia zależność współczyników (na znormalizowanych danych) od lambdy.






Jeżeli chodzi o selekcję cech za pomocą optymalnego lambda, odrzuciliśmy 10 z 32 cech. Cechy **odrzucone** przez takie lambda to:

- `White` - procent ludności białej.
- `Black` - procent ludności czarnej.
- `Pacific` - procent ludności z obszaru pacyficznego.
- `Construction` - procent ludności pracującej w budownictwie.
- `Production` - procent ludności pracującącej w branży produkcyjnej.
- `Drive` - procent ludności jeżdżącej do pracy samochodem.
- `WorkAtHome` - procent ludności pracującej w domu.
- `PrivateWork` - procent ludności pracującej w sektorze prywatnym.
- `FamilyWork` - procent ludności wykonującej prace domowe?.
- `Unemployment` - procent bezrobocia.