Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 219 · 2015

**Jacek Stelmach**

Uniwersytet Ekonomiczny w Katowicach
Wydział Zarządzania
Katedra Statystyki
Jacek.Stelmach@polwax.pl

# O WPŁYWIE WYBRANYCH METOD SELEKCJI NIELINIOWYCH ZMIENNYCH OBJAŚNIAJĄCYCH NA JAKOŚĆ MODELI REGRESYJNCH

**Streszczenie:** Najpopularniejsza parametryczna metoda najmniejszych kwadratów oraz jej rozszerzenia (regresja grzbietowa, metoda LASSO, metoda LARS, regresja BRIDGE) pozwalają na budowę addytywnych modeli liniowych. W rzeczywistości często mamy do czynienia z nieliniowymi zależnościami, a użyteczna informacja jest powtarzana w wielu zmiennych objaśniających. Bezkrytyczne wykorzystanie wszystkich takich dostępnych zmiennych może prowadzić do naruszenia założeń Gaussa-Markowa i najczęściej obniża jakość modeli regresyjnych. Znane metody selekcji pozwalają na wybór zmiennych, które wnoszą najwięcej użytecznej informacji, ograniczając jednocześnie zbędny szum. Opisany eksperyment weryfikuje metodą symulacji komputerowej jakość modeli regresyjnych otrzymanych za pomocą wybranych metod parametrycznych, dla których przeprowadzono selekcję predyktorów, wykorzystując: drzewa regresyjne, regresję grzbietową oraz algorytm genetyczny.

**Słowa kluczowe:** model regresyjny, selekcja predyktorów, algorytm genetyczny.

## Wprowadzenie

Selekcja zmiennych objaśnianych jest obok wyboru metody regresyjnej kluczową decyzją mającą wpływ zarówno na dopasowanie, jak i dokładność prognoz modeli regresyjnych. Zbyt mała, jak i zbyt duża ich liczba może negatywnie wpłynąć na jakość takich modeli. Stosując do oszacowania parametrów addytywnego modelu liniowego metodę najmniejszych kwadratów, najczęściej selekcję przeprowadza się metodami: wprowadzania, usuwania, eliminacji wstecznej oraz selekcji postępującej. Warto jednak rozważyć także inne metody selekcji, w tym metody heurystyczne (algorytm genetyczny).

## 1. Regresyjne liniowe metody parametryczne

Poza najbardziej popularną, klasyczną wielowymiarową metodą najmniejszych kwadratów, opracowane zostały jej uogólnienia, które umożliwiły budowę poprawnych modeli także wówczas, gdy naruszone są założenia Gaussa-Markowa (np. homoskedastyczność). Najbardziej znane rozszerzenia przedstawia tabela 1.

**Tabela 1.** Najbardziej znane rozszerzenia metody najmniejszych kwadratów

| Nazwa metody | Funkcja straty | Estymator parametrów | Właściwości |
|---|---|---|---|
| MNK – metoda najmniejszych kwadratów | $(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\mathbf{y} - \mathbf{X\beta})$ | $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$ | MNK-estymator jest *BLUE*, o ile spełnione są założenia Gaussa-Markowa. |
| UMNK – uogólniona MNK | $(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{y} - \mathbf{X\beta})$ | $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{y}$ | Rozszerzenie MNK, kiedy zachodzi autokorelacja lub heteroskedastyczność reszt. UMNK-estymator jest *BLUE*, jeśli $\boldsymbol{\Sigma}$ jest znana. |
| Ważona MNK | $(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}\mathbf{W}(\mathbf{y} - \mathbf{X\beta})$ | $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$ gdzie $\boldsymbol{W}$ – macierz wag | Lepsze właściwości estymatora niż dla MNK. Możliwe inne dobieranie wag (np. starsze dane – mniejsze wagi), które podniosą precyzję modelu. |
| Regresja grzbietowa | Dodanie pewnej stałej $\lambda$ w formie „kary" do funkcji straty: $(\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\mathbf{y} - \mathbf{X\beta}) + \lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta}$ | $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$ | Redukcja wariancji modelu kosztem jego obciążenia (mniejszy MSE estymatorów niż dla MNK). Umożliwia estymację nawet przy wysokiej współliniowości. |
| Metoda *LASSO* | Funkcja straty jak w MNK, dodane ograniczenia w formie parametru $t$ estymatorów: $\sum_{j=1}^{p}\lvert\beta_j \le t\rvert$ | Zaawansowane metody doboru parametru $t$ oraz estymacji $\widehat{\beta}_j$ | Eliminowane są zmienne, których estymatory mają małą wartość. W efekcie maleje błąd średniokwadratowy. |
| Regresja *Bridge* | Uogólnienie regresji grzbietowej, w której funkcja kary przyjmuje postać: $\lambda\sum_{i=1}^{p}\lvert\beta_j\rvert^{\gamma}$ dla $\gamma \ge 1$ | Zaawansowane metody doboru parametrów $\lambda$ i $\gamma$ oraz estymacji $\widehat{\beta}_j$ | Eliminowane są zmienne, których estymatory mają małą wartość. W efekcie maleje błąd średniokwadratowy. Działanie nie jest efektywne, jeśli występuje wiele estymatorów $\widehat{\beta}_j$ o małych wartościach. |
| Metoda *LARS (Least Angle Regression)* | Krokowe dołączenie zmiennych poprzez modyfikację estymatora $\hat{\beta}_j$ | Zmiana estymatora w kroku $k$: $\hat{\beta}_j(\alpha) = \hat{\beta}_j + \alpha\delta_k$, gdzie: $\delta_k = (\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{X}_{\boldsymbol{k}})^{-1}\boldsymbol{X}_{\boldsymbol{k}}^T\boldsymbol{e}_{\boldsymbol{k}}$; $e_k$ – bieżące reszty modelu; $\boldsymbol{X}_{\boldsymbol{k}}$ – macierz $\boldsymbol{X}$ dla dołączonych predyktorów | Mniejszy MSE modelu. Uproszczenie modelu przy dużej liczbie predyktorów (eliminacja zmiennych niewiele wnoszących do modelu). |

Źródło: Dittmann [2008], Efron i in. [2004]), Fox i Weisberg [2011], Fu [1998], Geladi i Kowalski [1986], Hastie i in. [2008], Hawkins [1994], Helland [1999], Hoerl i Kennard [1970], Hoskuldson [1998], Huber [1964], Huber i Ronchetti [2009], Maddala [2008], Tibshirani [1996], Wassermann [2006], Wilcox [2010].

W metodach tych może dojść do ograniczenia liczby predyktorów poprzez nałożenie celowych restrykcji na estymatory współczynników modelu lub funkcję straty.

## 2. Wybrane metody selekcji zmiennych objaśniających

Celem selekcji jest identyfikacja zmiennych, których eliminacja poprawia właściwości modelu regresyjnego. Jak podkreślają: Faraway [2002], Hastie i in. [2008] oraz Maddala [2008], nadmierna liczba predyktorów jest niekorzystna z uwagi na:

- ryzyko nadmiernej współliniowości predyktorów i związanych z tym problemów,
- wprowadzenie do modelu niepotrzebnej informacji (szumu) i niecelowej utraty stopni swobody, czego skutkiem jest zwiększona wariancja parametrów modelu (pomimo małego obciążenia),
- koszt przygotowania i pozyskania obserwacji rozbudowanych o nadmiarowe predyktory,
- trudności w interpretacji najbardziej znaczącego wpływu predyktorów na zmienną objaśnianą.

Najczęściej stosowane metody selekcji zmiennych objaśniających obejmują techniki:

- wprowadzania – wszystkie zmienne w określonym bloku są jednocześnie wprowadzane do modelu,
- usuwania – wszystkie zmienne w określonym bloku są jednocześnie usuwane z modelu,
- eliminacji wstecznej – po wprowadzeniu wszystkich zmiennych usuwana jest zmienna spełniająca kryteria usunięcia, aż do wyczerpania się zmiennych spełniających kryteria,
- selekcji postępującej – wprowadzanie do modelu kolejno zmiennych spełniających kryteria wprowadzenia, zaczynając od zmiennej, która w najwyższym stopniu spełnia przyjęte kryterium, aż do wyczerpania się zmiennych spełniających kryteria.

Znana literatura podaje wiele kryteriów wyboru, wybrane kryteria podano w tabeli 2 (gdzie $N$ oznacza liczbę obserwacji, zaś $p$ – liczbę parametrów modelu). O ile spełnione są założenia Gaussa-Markowa, największą funkcją wiarygodności modelu liniowego jest wyrażenie: $L = \frac{RSS}{N}$. Do często stosowanych metod ograniczających ich liczbę należy także analiza głównych składowych (PCA – *Principal Component Analysis*), w której po wyznaczeniu nowych zmiennych (jako kombinacji liniowej zmiennych objaśniających) następuje ich ograniczenie zgodnie

z pewnym wybranym kryterium. Najczęściej wykorzystywane są: kryterium osypiska, kryterium Kaisera, odsetek wyjaśnionej wariancji [Jackson, 1991]. Metoda ta wprawdzie zmniejsza liczbę predyktorów w modelu regresyjnym, ale nie eliminuje wpływu zmiennych niosących wyłącznie szum. Ułomnością tej metody jest ponadto utrudniona interpretacja współczynników modelu (współczynniki odnoszące się do oryginalnych predyktorów występują w postaci uwikłanej).

**Tabela 2.** Wybrane kryteria wyboru zmiennych objaśniających

| Kryterium | Wyrażenie |
|---|---|
| Max współczynnika determinacji $R^2$ | $R^2 = 1 - \frac{RSS}{SST}$ |
| Min współczynnika Hockinga $S_p$ | $S_p = \frac{RSS}{N - p}$ |
| Min współczynnika Mallowsa $C_p$ | $C_p = \frac{RSS}{MSE} - (N - 2p)$ |
| Min kryterium Akaike AIC | $AIC = 2p - 2 \times \ln(L)$ |
| Min kryterium Bayesowskiego BIC | $BIC = p \times ln(N) - 2 \times \ln(L)$ |
| Max statystyki F | $F = \frac{MSR}{MSE}$ |

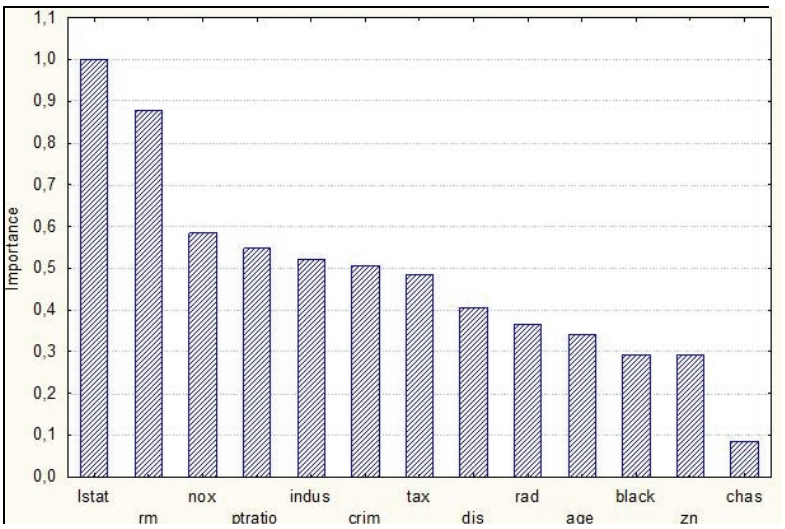
Źródło: Maddala [2008], Rencher [2002].

Do określenia optymalnego zestawu zmiennych objaśniających można także wykorzystać pojemność integralną kombinacji nośników informacji opracowaną przez Zdzisława Hellwiga [Barczak i Biolik, 2003], w której dąży się do jak największej korelacji predyktorów ze zmienną objaśnianą przy jednoczesnej słabej korelacji pomiędzy tymi predyktorami.

Jak wcześniej podkreślono, właściwy dobór zmiennych ma kluczowe znaczenie dla stopnia dopasowaniu modeli regresyjnych i dokładności prognoz. Nie mniej ważna jest identyfikacja, czy zależność wybranych predyktorów na zmienną objaśnianą jest liniowa. Przy dużych nieliniowościach jakość otrzymanych modeli pomimo prawidłowego doboru zmiennych objaśniających może być niezadowalająca. W takich przypadkach można posiłkować się wprowadzeniem dodatkowych predyktorów otrzymanych z oryginalnych zmiennych za pomocą odpowiednio dobranych funkcji nieliniowych. Jeśli taka identyfikacja jest utrudniona, można wprowadzić dodatkowe zmienne, wykorzystując najbardziej popularne funkcje nieliniowe, i przeprowadzić proces wyboru, który powinien wyłonić tylko te nieliniowe zmienne, które wnoszą do modelu użyteczną informację. Niestety, jeśli stosowanym kryterium jest na przykład wynik testu *t* istotności współczynników modelu, powyższe techniki nie mogą być wykorzystane do budowy modeli nieparametrycznych. Poniżej przedstawiono zatem takie techniki, które można wykorzystać także w tworzeniu modeli nieparametrycznych:

- przegląd widma zmiennych (analiza Fouriera), Mix [1995] wskazuje, że zmienne będące szumem (które nie niosą informacji) mają znacznie szersze widmo, czyli mają znacznie więcej istotnych składowych widmowych od zmiennych niosących informację,
- wykorzystanie rangowania zmiennych objaśniających w drzewach regresyjnych *CART* (rys. 1), umożliwiające usunięcie zmiennych, dla których ich 'ważność' jest wyraźnie mniejsza od pozostałych zmiennych,
- wykorzystanie algorytmów genetycznych, umożliwiających znalezienie suboptymalnego zestawu bez konieczności przeszukania wszystkich kombinacji predyktorów zgodnie z pewnym wybranym kryterium (np. zgodnie z tabelą 2),
- wybór dodatkowych nieliniowych funkcji zmiennych objaśniających w miejsce oryginalnych zmiennych, z wykorzystaniem także algorytmów genetycznych.

Skuteczną metodą weryfikacji, czy nie doszło do przeuczenia modelu, może być sprawdzian krzyżowy (*cross-validation)*, w którym dokonuje się podziału na *K* rozłącznych podzbiorów, tworzących następnie zbiór uczący i zbiór testowy.




**Rys. 1.** Przykład określenia ważności predyktorów uzyskanych za pomocą drzew regresyjnych

## 3. Prezentacja eksperymentu

W eksperymencie dokonano porównania wybranych metod selekcji zmiennych objaśniających, wykorzystując symulowane zbiory danych:

1. Zbiory zalecane przez Friedmana [1991], symulujące szumy i zmienne elektryczne pewnego obwodu elektronicznego:

– *Friedman_1*, zbiór danych stanowi 10 zmiennych objaśniających $x_1, \ldots, x_{10}$ o rozkładzie równomiernym w przedziale [0, 1], zmienna objaśniana wykorzystująca wyłącznie zmienne $x_1, \ldots, x_5$ opisana jest wzorem:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + e_1\,,$$

gdzie: $e_1 \sim N(0,1)$.

– *Friedman_2*, zbiór danych stanowią 4 zmienne objaśniające o rozkładzie równomiernym w przedziałach:

$$0 \le x_1 \le 100,$$
$$40\pi \le x_2 \le 560\pi,$$
$$0 \le x_3 \le 1,$$
$$1 \le x_4 \le 11,$$

zaś zmienna objaśniana opisana jest wzorem:

$$y = \left[x_1^2 + \left(x_2 x_3 - \frac{1}{x_2 x_4}\right)^2\right]^{0.5} + e_2\,,$$

gdzie: $e_2 \sim N(0,9)$.

– *Friedman_3*, zbiór danych stanowią 4 zmienne objaśniające o rozkładzie jak dla zbioru „*Friedman_2"*, zmienna objaśniana opisana jest wzorem:

$$y = tan^{-1}\left(\frac{x_2 x_3 - \frac{1}{x_2 x_4}}{x_1}\right) + e_3,$$

gdzie: $e_3 \sim N(0,1)$.

2. Zbiory symulowane o rozkładzie normalnym wielowymiarowym:

– *Zestaw1*, zbiór z nieskorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$~N(0,1), zmienna objaśniana obliczona została zgodnie ze wzorem:

$$y = \sum_{i=1}^{10} \beta_i\, x_i + \frac{e}{10}\,,$$

gdzie: $\beta_1, \ldots, \beta_{10}$ wylosowano z przedziału (-1, 1), błąd losowy $e \sim$ N(0,1).

– *Zestaw2*, zbiór z nieskorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$~N(0,1), zmienna objaśniana opisana jest wzorem:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_4 + \frac{e}{10}\,,$$

gdzie: $\beta_i$, $e$ – jak wyżej.

– *Zestaw3*, zbiór ze skorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$~N(0,1), zgodnie z macierzą kowariancji przedstawioną w tabeli 3, zmienna objaśniana opisana jest wzorem:

$$y = \sum_{i=1}^{10} \beta_i x_i + \frac{e}{10} ,$$

gdzie: $\beta_i$, $e$ – jak wyżej.

– *Zestaw4*, zbiór ze skorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$~N(0,1), zgodnie z macierzą kowariancji przedstawioną w tabeli 3, zmienna objaśniana opisana jest wzorem:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_4 + \frac{e}{10} ,$$

gdzie: $\beta_i$, $e$ – jak wyżej.

– *Zestaw5*, zbiór ze skorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$~N(0,1), zgodnie z macierzą kowariancji przedstawioną w tabeli 4 (większe zależności niż dla *„Zestawu 3”*), zmienna objaśniana opisana jest wzorem:

$$y = \sum_{i=1}^{10} \beta_i x_i + \frac{e}{10} ,$$

gdzie: $\beta_i$, $e$ – jak wyżej.

Dodatkowo dla 5% obserwacji wartość zmiennych objaśnianych zastąpiono losowo wartością pięciokrotnie większą (symulacja dodatkowych zakłóceń).

**Tabela 3.** Macierz kowariancji *„Zestawu3”* oraz *„Zestawu4”*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0,4 | 1 | 0,4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0,4 | 1 | 0,4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0,4 | 1 | 0,4 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0,4 | 1 | 0,4 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0,4 | 1 | 0,4 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0,4 | 1 | 0,4 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0,4 | 1 | 0,4 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,4 | 1 | 0,4 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,4 | 1 |

– *Zestaw6,* zbiór ze skorelowanymi zmiennymi objaśniającymi $x_1, \ldots, x_{10}$ ~ N(0,1), zgodnie z macierzą kowariancji przedstawioną w tabeli 4 (większe zależności niż dla *„Zestawu 3”*), zmienna objaśniana opisana jest wzorem:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_4 + \frac{e}{10} \ ,$$

gdzie: $\beta_i$, $e$ – jak wyżej.

Dodatkowo dla 5% obserwacji wartość zmiennych objaśnianych zastąpiono losowo wartością pięciokrotnie większą (symulacja dodatkowych zakłóceń).

**Tabela 4.** Macierz kowariancji *„Zestawu5”* oraz *„Zestawu6”*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,7 | 0,7 | 0,7 | 0,7 | 0,7 | 0,7 | 0,7 | 0,7 | 0,9 |
| 0,7 | 1 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,9 |
| 0,7 | 0,5 | 1 | 05 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,6 |
| 0,7 | 0,5 | 0,5 | 1 | 0,5 | 0,5 | 0,5 | 0,5 | 0,6 | 0,7 |
| 0,7 | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,5 | 0,5 | 0,5 | 0,6 |
| 0,7 | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,5 | 0,5 | 0,6 |
| 0,7 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,5 | 0,6 |
| 0,7 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 0,5 | 0,6 |
| 0,7 | 0,5 | 0,5 | 0,6 | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 0,6 |
| 0,9 | 0,9 | 0,6 | 0,7 | 0,6 | 0,6 | 0,6 | 0,6 | 0,6 | 1 |

Z uwagi na to, że większość zależności pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi (poza *„Zestawem1”*, *„Zestawem3”* oraz *„Zestawem5”*) ma charakter nieliniowy, zdecydowano się na przeprowadzenie badań dla oryginalnych zestawów zmiennych objaśniających (typ *short*) oraz zmiennych uzupełnionych o ich funkcje nieliniowe – kwadrat, arcus tangens, logarytm modułu, odwrotność (typ *long*), co znacznie zwiększyło liczbę predyktorów. Takie poszerzenie zmiennych objaśniających powinno poprawić dopasowanie modeli regresyjnych oraz dokładność prognoz. Ponieważ znane są wzory opisujące zależności, można będzie ocenić trafność selekcji zmiennych wybranymi metodami.

Zdecydowano się na selekcję zmiennych, wybierając spoza wyżej rozważanych metod:

- eliminację wsteczną regresji liniowej zgodnie z kryterium informacyjnym *AIC*,
- eliminację wsteczną regresji liniowej zgodnie z współczynnikiem *Cp* Mallowsa,
- selekcję metodą regresji grzbietowej,
- algorytm genetyczny zgodnie z kryterium informacyjnym *AIC*,
- algorytm genetyczny zgodnie z kryterium informacyjnym *BIC*,
- algorytm genetyczny zgodnie z wielkością integralnej pojemności kombinacji nośników informacji (kryterium Hellwiga),
- algorytm genetyczny zgodnie z współczynnikiem determinacji $R^2$,
- algorytm genetyczny zgodnie z pierwiastkiem błędu średniokwadratowego *RMSE,*
- „ważność zmiennych” określoną za pomocą drzew regresyjnych.

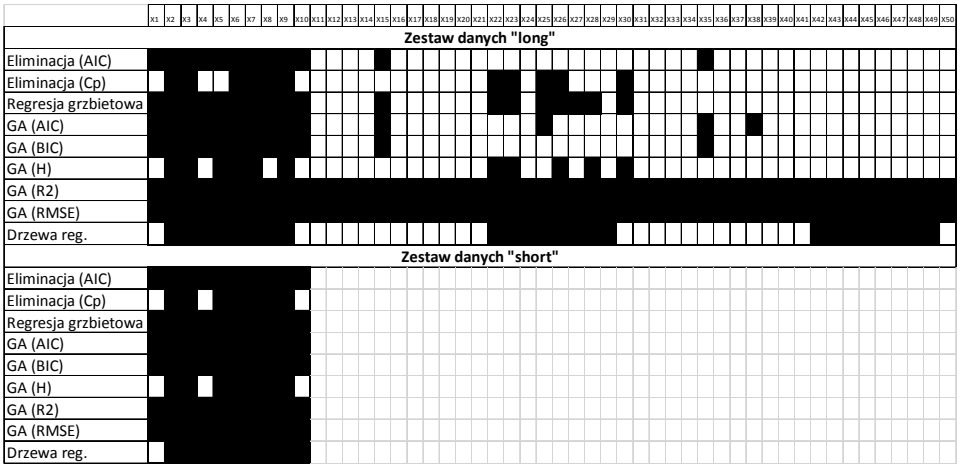
Selekcję zmiennych objaśniających za pomocą algorytmów genetycznych przeprowadzono, wykonując 100 iteracji, wybierając większościowy zestaw zmiennych objaśniających. Następnie porównano modele regresyjne zbudowane

w oparciu o wybrane zmienne objaśniające z modelami wykorzystującymi zmienne po transformacji PCA (dobierając zmienne objaśniające kryteriami: Kaisera i osypiska) za pomocą metod parametrycznych:
- klasycznej MNK,
- regresji LASSO,
- regresji LARS,
- regresji grzbietowej,
- regresji BRIDGE.

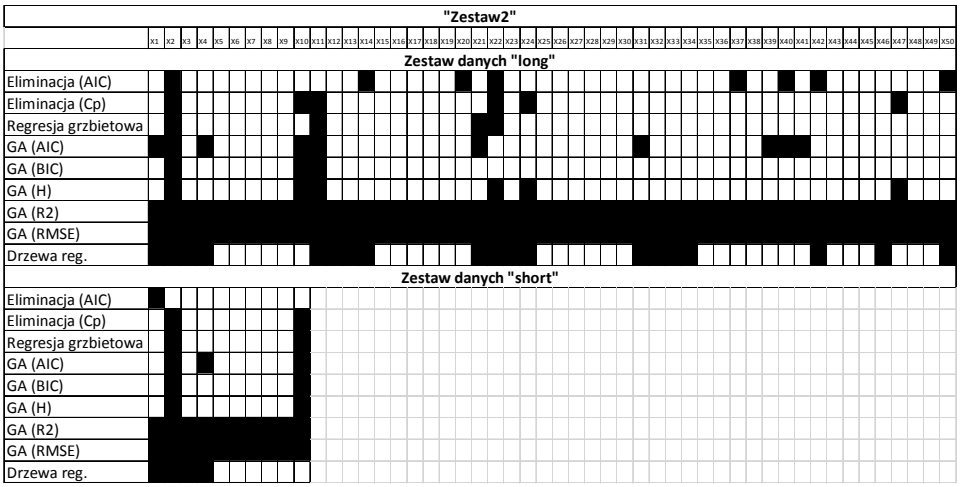
Wszystkie obliczenia przeprowadzono za pomocą programu R 3.1.0, wykorzystując następujące pakiety: *BayesBridge, GA, lars, leaps, MASS* oraz *ridge.* Pakiet *GA* pozwala na kompletne wykorzystanie algorytmu genetycznego. Umożliwia niezależny wybór metod selekcji, krzyżowania i mutacji. W eksperymencie wykorzystano ten pakiet z przyjętymi domyślnymi metodami, które można odczytać za pomocą polecenia gaControl. Fragment skryptu przedstawiający wykorzystanie algorytmu zamieszczono w załączniku.

## 4. Omówienie wyników

Najbardziej interesujące wyniki selekcji zmiennych objaśniających dla badanych zbiorów danych przedstawiają rys. 2-9, określające włączenie zmiennej objaśniającej do modelu przez zaczernienie odpowiadającego jej pola. Można zauważyć, że metoda wykorzystująca algorytm genetyczny z funkcjami przystosowania: współczynnik determinacji $R^2$ oraz miara *RMSE* praktycznie dopuszczają wszystkie zmienne objaśniające (pomimo że w *„Zestawie1”* występują wyłącznie zależności liniowe). Stwarza to ogromne ryzyko nadmiernego dopasowania modelu i utracenia właściwości generalizacji, a w konsekwencji gorszych właściwości prognostycznych.

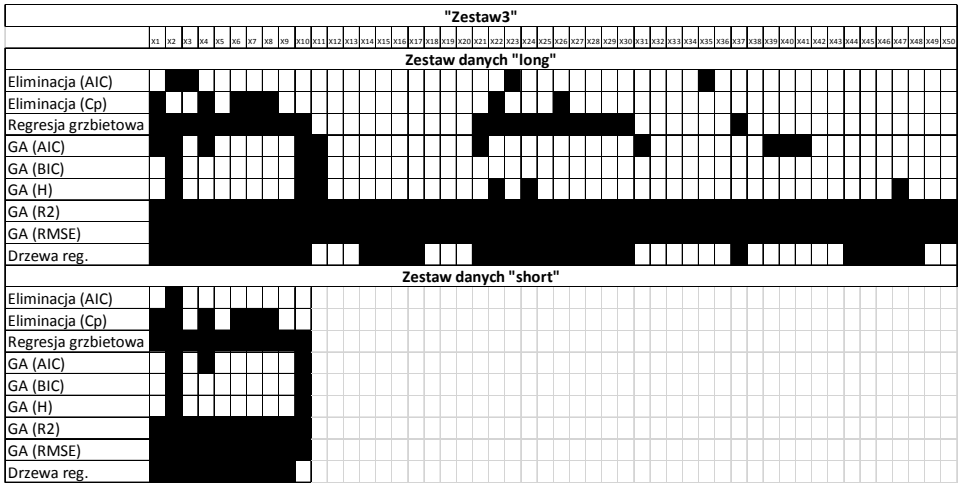



**Rys. 2.** Wyniki selekcji predyktorów *„Zestawu1”*

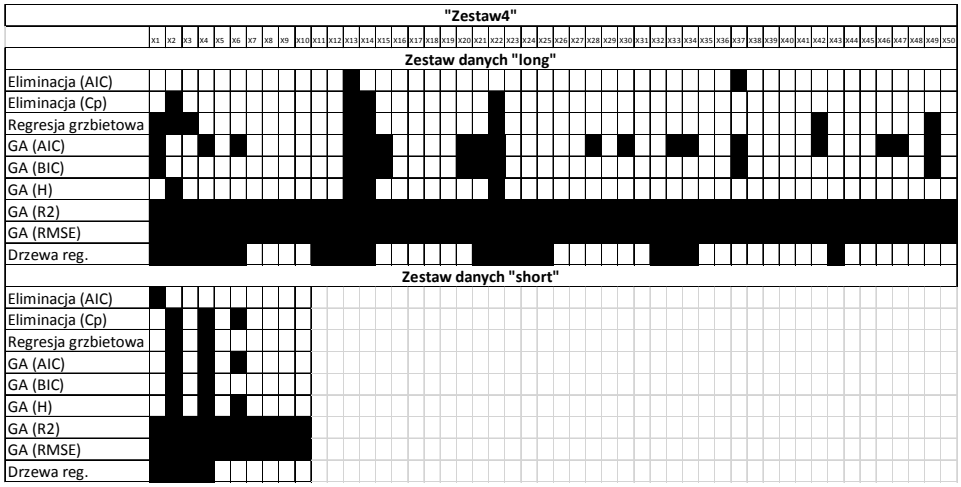



**Rys. 3.** Wyniki selekcji predyktorów *„Zestawu2”*

Bardzo dobrą selekcję zmiennych dla *„Zestawu1”* przeprowadziły wszystkie metody (poza algorytmem genetycznym z $R^2$ i *RMSE* jako funkcjami przystosowania), zapewne z uwagi na liniowe zależności tego zbioru danych. Niestety, wtrącone w także liniowym *„Zestawie5”* zakłócenia doprowadziły do niesłusznej eliminacji części oryginalnych predyktorów. Z kolei w badaniach zbiorów *„Zestaw2”*, *„Zestaw4”* oraz *„Zestaw6”*, dla których dominowały zależności nieliniowe, znacznie częściej wybierane także były nieliniowe funkcje oryginalnych zmiennych objaśniających.

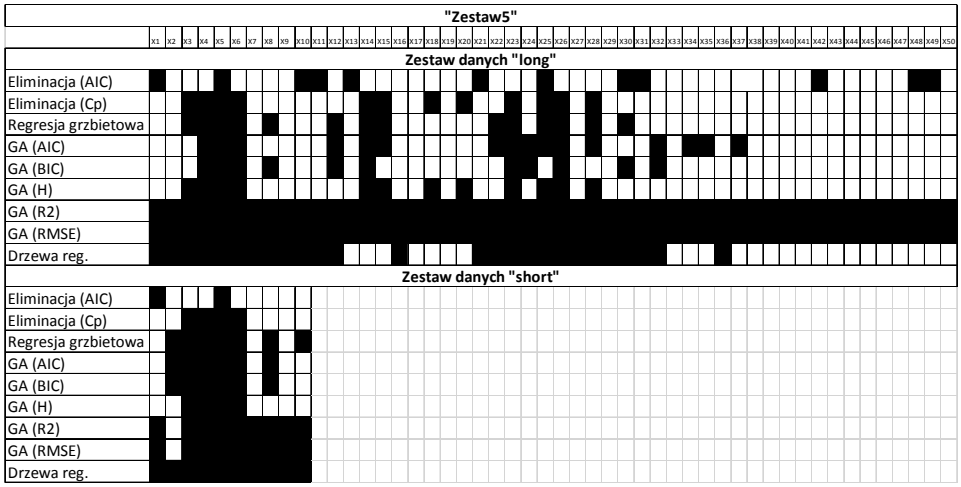



**Rys. 4.** Wyniki selekcji predyktorów *„Zestawu3”*

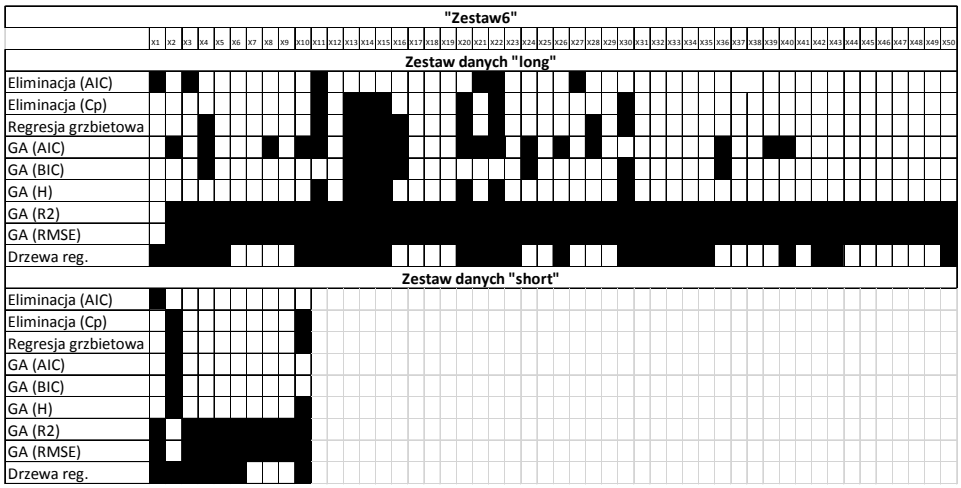



**Rys. 5.** Wyniki selekcji predyktorów *„Zestawu4”*

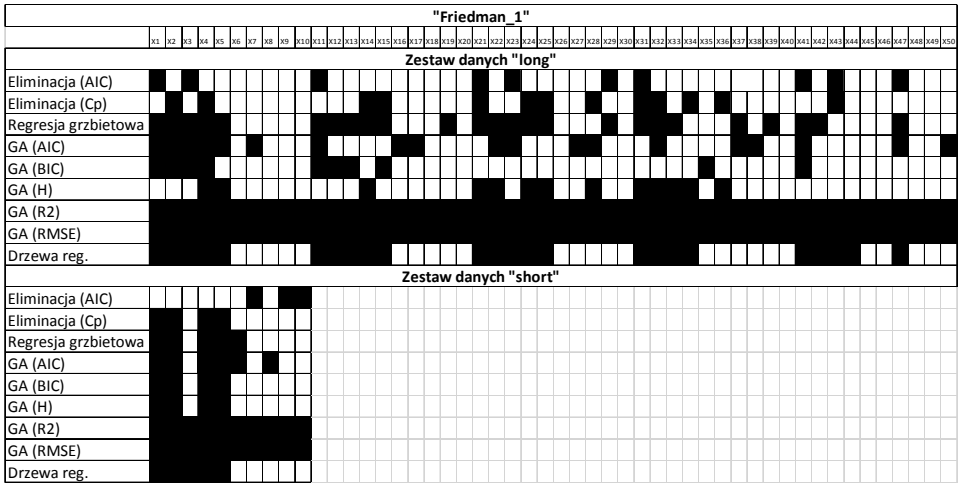
Z kolei kryterium integralnej pojemności kombinacji nośników informacji (Hellwiga) odrzuca zbyt dużo zmiennych. Twierdzenie to popierają średnie wartości miar dopasowania modelu regresji liniowej, przedstawione przykładowo w tabelach 5 i 6.



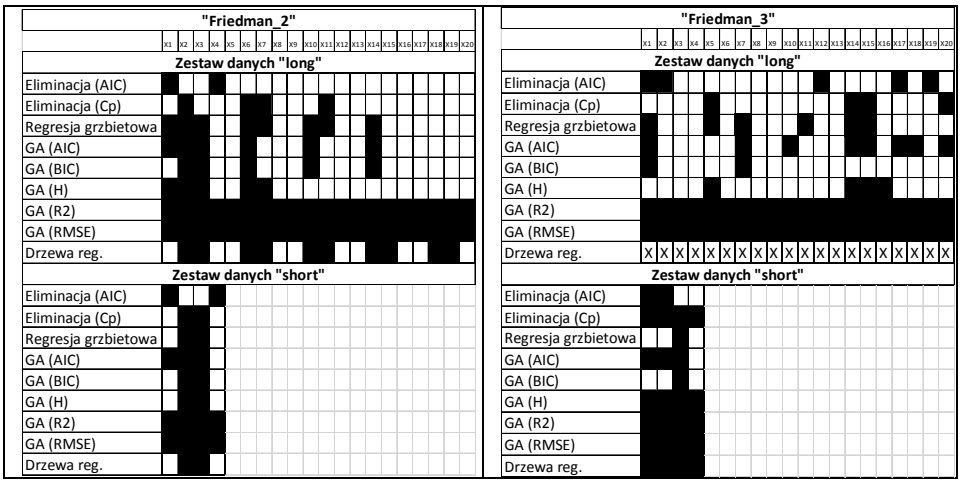

**Rys. 6.** Wyniki selekcji predyktorów *„Zestawu5”*




**Rys. 7.** Wyniki selekcji predyktorów *„Zestawu6”*




**Rys. 8.** Wyniki selekcji predyktorów *„Friedman_1”*




**Rys. 9.** Wyniki selekcji predyktorów *„Friedman_2”* oraz *„Friedman_3”*

**Tabela 5.** Miary dopasowania zmiennych objaśniających wybranych za pomocą algorytmu genetycznego dla *„Zestawu1”*

| Kryterium algorytmu genetycznego | Zestaw *long* poszerzony o funkcje nieliniowe | | | | Zestaw *short* – zmienne bez modyfikacji | | | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | RMSE | $R^2$ | AIC | BIC | RMSE | $R^2$ |
| minimum AIC | -320,8 | -268,1 | 0,,1156 | 0.9982 | -310,8 | -271,2 | 0,1210 | 0,9981 |
| minimum BIC | -318,3 | -272,1 | 0,1176 | 0,9982 | -310,8 | -271,2 | 0,1210 | 0,9981 |
| maksimum pojemności Hellwiga | 291,1 | 334,0 | 0,5421 | 0,9610 | 382,3 | 411,9 | 0,6946 | 0,9360 |
| maksimum współczynnika $R^2$ | -262,3 | -90,78 | 0,1118 | 0,9983 | -310,8 | -271,2 | 0,1210 | 0,9981 |
| minimum błędu RMSE | -262,3 | -90,78 | 0,1118 | 0,9983 | -310,8 | -271,2 | 0,1210 | 0,9981 |

**Tabela 6.** Miary dopasowania zmiennych objaśniających wybranych za pomocą algorytmu genetycznego dla *„Friedman_1”*

| Kryterium algorytmu genetycznego | Zestaw *long* poszerzony o funkcje nieliniowe | | | | Zestaw *short* – zmienne bez modyfikacji | | | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | RMSE | $R^2$ | AIC | BIC | RMSE | $R^2$ |
| minimum AIC | 719,4 | 788,6 | 1,519 | 0,925 | 945,9 | 972,3 | 2,857 | 0,733 |
| minimum BIC | 735,8 | 775,4 | 1,656 | 0,910 | 948,3 | 968,1 | 2,903 | 0,725 |
| maksimum pojemności Hellwiga | 926,7 | 976,2 | 2,629 | 0,774 | 948,3 | 968,1 | 2,903, | 0,725 |
| maksimum współczynnika $R^2$ | 749,4 | 921,0 | 1,403 | 0,936 | 951,2 | 990,8 | 2,838 | 0,737 |
| minimum błędu RMSE | 749,4 | 921,0 | 1,403 | 0,936 | 951,2 | 990,8 | 2,838 | 0,737 |

Wyniki uzupełniają miary obliczone dla modeli opartych na zmiennych objaśniających poddanych transformacji PCA, a następnie ograniczonych zgodnie z kryteriami Kaisera i osypiska. Stopień redukcji zmiennych przedstawia tabela 7.

**Tabela 7.** Zmienne po transformacji PCA i selekcji za pomocą wybranych kryteriów

| Nazwa zestawu danych | Zmienne $X\_PCA_i$ |
|---|---|
| *„Friedman_1”* | $X\_PCA_1$ - $X\_PCA_4$ |
| *„Friedman_2”* | $X\_PCA_1$ - $X\_PCA_2$ |
| *„Friedman_3”* | $X\_PCA_1$ - $X\_PCA_2$ |
| *„Zestaw1”* | $X\_PCA_1$ - $X\_PCA_6$ |
| *„Zestaw2”* | $X\_PCA_1$ - $X\_PCA_6$ |
| *„Zestaw3”* | $X\_PCA_1$ - $X\_PCA_5$ |
| *„Zestaw4”* | $X\_PCA_1$ - $X\_PCA_5$ |
| *„Zestaw5”* | $X\_PCA_1$ - $X\_PCA_3$ |
| *„Zestaw6”* | $X\_PCA_1$ - $X\_PCA_3$ |

Na uwagę zasługują także wyniki selekcji za pomocą regresji grzbietowej i drzew regresyjnych. Wybrane zmienne są do siebie bardzo zbliżone i w większości zgodne z oczekiwaniami wynikającymi ze znajomości wzorów, na podstawie których wygenerowano badane zbiory danych.

Końcową ocenę przeprowadzono, dokonując większościowego wyboru predyktorów metodami: regresji grzbietowej (także Cule i De Iorio [2012] potwierdzili przewagę wyboru zmiennych objaśniających za pomocą regresji grzbietowej w porównaniu z klasyczną MNK oraz metodą LASSO), algorytmem genetycznym z funkcją przystosowania *AIC* oraz oceniając „ważność” zmiennych za pomocą drzew regresyjnych.

Ostateczne wyniki eksperymentu stanowi ocena modeli regresyjnych obejmująca zarówno dopasowanie (*RMSE)*, jak i dokładność prognoz (*MAE*), przeprowadzona za pomocą sprawdzianu krzyżowego. Wyniki badań umieszczono w tabelach 8-10.

**Tabela 8.** Średnie wartości miar *RMSE* oraz *MAE* modelu obliczonego dla zbioru *„Zestaw1”*

| Zbiór danych | Metody regresji parametrycznej | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MNK | | LASSO | | LARS | | r. grzbietowa | | BRIDGE | |
| | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| *short* | 0,105 | 0,079 | 0,106 | 0,079 | 0,106 | 0,079 | 1,097 | 0,983 | 0,106 | 0,079 |
| *long* | 0,103 | 0,077 | 0,101 | 0,094 | 0,101 | 0,094 | 0,116 | 0,092 | 0,099 | 0,090 |
| PCA | 0,786 | 0,740 | 0,789 | 0,735 | 0,789 | 0,735 | 0,746 | 1,606 | 0,789 | 0,734 |

**Tabela 9.** Średnie wartości miar *RMSE* oraz *MAE* modelu obliczonego dla zbioru *„Zestaw2”*

| Zbiór danych | Metody regresji parametrycznej | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MNK | | LASSO | | LARS | | r. grzbietowa | | BRIDGE | |
| | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| *short* | 0,953 | 0,769 | 0,929 | 0,823 | 0,929 | 0,823 | 0,952 | 0,760 | 0,932 | 0,798 |
| *long* | 0,881 | 0,513 | 1,114 | 0,856 | 1,114 | 0,856 | 0,895 | 0,453 | 0,912 | 0,434 |
| PCA | 1,045 | 0,738 | 1,042 | 0,705 | 1,042 | 0,705 | 1,045 | 0,724 | 1,044 | 0,720 |

**Tabela 10.** Średnie wartości miar *RMSE* oraz *MAE* modelu obliczonego dla zbioru *„Friedman_1”*

| Zbiór danych | Metody regresji parametrycznej | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MNK | | LASSO | | LARS | | r. grzbietowa | | BRIDGE | |
| | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE |
| *short* | 2,956 | 3,011 | 2,788 | 2,771 | 2,788 | 2,771 | 2,958 | 3,012 | 2,901 | 2,998 |
| *long* | 2,482 | 2,468 | 2,497 | 2,443 | 2,497 | 2,443 | 2,658 | 2,530 | 2,453 | 2,429 |
| PCA | 3,748 | 3,384 | 6,737 | 4,836 | 6,737 | 4,836 | 3,749 | 3,386 | 4,041 | 3,476 |

We wszystkich przypadkach (poza *„Zestawem1”*) można zauważyć przewagę modeli budowanych w oparciu o zbiory danych rozszerzone o funkcje nieliniowe predyktorów. Jest to intuicyjnie zrozumiałe – skoro zależności pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi (poza *„Zestawem1”* i *„Zestawem5”*) są nieliniowe, znacznie dokładniejsze modele regresyjne można zbudować, wykorzystując także zmienne poddane nieliniowym przekształceniom. Selekcja zmiennych spowodowała, że różnice pomiędzy błędami modeli otrzymanych za pomocą klasycznej MNK oraz modeli otrzymanych przez rozszerzenia tej metody zmniejszyły się. Wydaje się, że wstępnie dokonana selekcja predyktorów zredukowała korzystny wpływ restrykcji nakładanych na estymatory parametrów w postaci funkcji straty. Największe wartości błędów dopasowania i prognoz odnoszą się do modeli uzyskanych dla danych poddanych transformacji PCA. Takie zachowanie może być efektem dobrze przeprowadzonej selekcji zmiennych objaśniających, która w większym stopniu zachowała użyteczną informację, niż poprzez selekcję zmiennych po transformacji PCA.

## Podsumowanie

Selekcja zmiennych objaśniających w bezpośredni sposób wpływa na jakość budowanych modeli regresyjnych. Zarówno zbyt mała, jak i zbyt duża ich

ilość prowadzi do pogorszenia właściwości prognostycznych modeli, wynikających bądź ze zbyt małej ilości dostarczonej informacji, bądź z przeuczenia modelu. Zaproponowane metody selekcji predyktorów mogą być alternatywne dla powszechnie stosowanych metod selekcji. Pozwalają one na wybór właściwych zmiennych, nawet jeśli ich liczba jest dość znaczna. Należy podkreślić, że wyniki badań potwierdziły celowość rozszerzenia oryginalnych predyktorów o ich funkcje nieliniowe, jeśli zachodzi podejrzenie nieliniowych związków pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi (nierzadko występujących w rzeczywistości). Efektem jest jednak znaczne zwiększenie liczby zmiennych objaśniających – co tym bardziej wymaga przemyślanej i dobrze przeprowadzonej selekcji, także proponowanymi metodami, których efektywność potwierdziły wyniki przeprowadzonych symulacji.

*Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/03/B/HS4/05630.*

## Literatura

Barczak A., Biolik J. (2003), *Podstawy ekonometrii*, Wydawnictwo AE, Katowice.

Cule E., De Iorio M. (2012), *A semi-automatic method to guide the choice of ridge parameter in ridge regression*, arXiv:1205.0686v1 [stat.AP].

Dittmann P. (2008), *Prognozowanie w przedsiębiorstwie*, Oficyna a Wolters Kluwer business, Kraków.

Efron B., Hastie T., Johnstone I., Tibshirani R. (2004), *Least Angle Regression*, „The Annals of Statistics, Vol. 32.

Faraway J.J. (2002), *Practical Regression and Anova using R*, http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf (dostęp: 12.10.2012).

Fox J., Weisberg S. (2011), *An R Companion to Applied Regression*, Sage, Thousand Oaks CA.

Friedman J.H. (1991), *Multivariate Adaptive Regression Splines*, „Annals of Statistics", Vol. 19(1).

Fu W.J. (1998), *Penalized Regressions: The Bridge Versus the Lasso*, „Journal of Computational and Graphical Statistics", Vol. 7.

Geladi P., Kowalski B. (1986), *Partial least square regression: A tutorial*, „Analytica Chemica Acta", Vol. 35.

Hastie T., Tibshirani R., Friedman J. (2008), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Science+Business Media B.V., New York.

Hawkins D.M. (1994), *The feasible solution algorithm for least trimmed squares regression*, „Computational Statistics & Data Analysis", Vol. 17.

Helland I.S. (1999), *Some theoretical aspects of partial least squares regression*, „Chemometrics and Intelligent Laboratory Systems", Vol. 58.

Hoerl A.E., Kennard R.W. (1970), *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, „Technometrics", Vol. 12.

Hoskuldson A. (1988), *Pls regression methods*, „Journal of Chemometrics", Vol. 2.

Huber P.J. (1964), *Robust estimation of a location parameter*, „The Annals of Mathematical Statistics", Vol. 35(1).

Huber P.J., Ronchetti E.M. (2009), *Robust Statistics*, A John Wiley & Sons, Inc. Publication, New Jersey.

Jackson E.J. (1991), *A User's Guide to Principal Components*, A John Wiley & Sons, Inc. Publication, New Jersey.

Maddala G.S. (2008), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.

Mix D.F. (1995), *Random Signal Processing*, Prentice Hall Inc.

Rencher A.C. (2002), *Methods of Multivariate Analysis*, A John Wiley & Sons, Inc. Publication, New Jersey.

Tibshirani R. (1996), *Regression Shrinkage and Selection via the Lasso*, „Journal of the Royal Statistical Society", Vol. 58.

Wasserman L. (2006), *All of nonparametric statistics*, Springer Science+Business Media B.V., New York.

Wilcox R.R. (2010), *Fundamentals of Modern Statistical Methods*, Springer Science+Business Media B.V., New York.

## ON THE IMPACT OF SOME METHODS OF SELECTION NONLINEAR VARIABLES ON QUALITY OF REGRESSION MODELS

**Summary:** The most common parametric Ordinary Least Squares Method and its extension (ridge regression, LASSO and LARS methods, BRIDGE regression) allow to build additive linear models. In reality, we often have to deal with non-linear dependencies, and useful information is repeated in a number of explanatory variables. Use of all available variables can lead to violations of Gauss-Markow assumptions and frequently reduces the quality of regression models. Known methods of selection allow to select the variables that contribute the most useful information, reducing unnecessary noise. Described experiment verifies, by computer simulation, quality of regression models obtained using selected parametric methods, for which the selection was carried out using: regression trees, ridge regression and genetic algorithm.

**Keywords:** regression model, selection of predictors, genetic algorithm.

## Załącznik

Poniżej przedstawiono fragment skryptu wykorzystywanego w prezentowanym eksperymencie, obejmujący stosowanie algorytmu genetycznego.

```
library(GA)

# Funkcje jakosci genomu, zaleznie od stosowanego przypadku

fitness1<-function(string)
{
 inc<-which(string==1)
 X<-cbind(1, (x[,inc]))
 mod<-lm.fit(X,y)
 class(mod)<-"lm"
 if (case=="AIC")
 return(-AIC(mod))
 if (case=="BIC")
 return(-BIC(mod))
 if (case=="SE")
 return(-sqrt(sum((mod$residuals)^2)/(dat_no-n_pred)))
 if (case=="R2")
 return(1-(sum(mod$residuals^2)/sum((y-mean(y))^2)))
 if (case=="H")
 return(Hellwig(X,y))
}
attach(dat1)
 mods=lm(model_s)
 xs=model.matrix(mods)[,-1]
 ys=model.response(model.frame(mods))
 case="AIC"
 GAs<-ga("binary", fitness=fitness1s, nBits=pred_no, na-
mes=colnames(xs),monitor=monitor,
 run=runit, maxiter=maxit)
 w=GAs@solution
 if (nrow(w)>1) w=w[1,]
 inc<-which(w==1)
 Xs<-cbind(1, xs[,inc])
 mods<-lm.fit(Xs, ys)
 class(mods)<-"lm"
 results_sAIC[MC,1]=AIC(mods)
 results_sAIC[MC,2]=BIC(mods)
 results_sAIC[MC,3]=sqrt(sum((mods$residuals)^2)/(dat_no-n_pred))
 results_sAIC[MC,4]=1-(sum(mods$residuals^2)/sum((ys-mean(ys))^2))
 results_sAIC[MC,5:(4+pred_no)]=w
```