

Joanna Trzęsiok

Akademia Ekonomiczna w Katowicach

**WYKORZYSTANIE TRANSFORMACJI ZMIENNYCH
W BUDOWIE MODELU REGRESYJNEGO****1. Wstęp**

Celem analizy regresji jest znalezienie z wykorzystaniem zbioru uczącego:

$$U = \{(\mathbf{x}_j, y_j) : j = 1, 2, \dots, N\},$$

funkcji g opisującej zależność zmiennej objaśnianej Y od zmiennych objaśniających $X_1, X_2, \dots, X_p$:

$$Y = g(X_1, X_2, \dots, X_p) + \varepsilon, \quad (1)$$

gdzie ε jest składnikiem losowym.

Jeśli założymy, że badana zależność ma charakter liniowy, uzyskujemy model:

$$Y = \sum_{i=1}^p \alpha_i X_i + \varepsilon. \quad (2)$$

Model regresyjny postaci (2) w wielu przypadkach charakteryzuje się słabym dopasowaniem do danych i dużymi błędami prognoz. Jednym ze sposobów uzyskania bardziej precyzyjnego modelu jest zastosowanie transformacji zmiennych:

$$\theta(Y) = \sum_{i=1}^p f_i(X_i) + \varepsilon. \quad (3)$$

Wykorzystanie w modelu regresyjnym odpowiednich przekształceń zmiennych jest bardzo popularne i powszechnie stosowane. Najczęściej jednak w zagadnieniu takim mamy do czynienia z klasycznymi metodami regresji, w których do transformacji zmiennych stosowane są np. funkcje logarytmiczne lub wielomianowe.

Poniższa praca ma na celu przedstawienie dwóch nieklasycznych metod regresji – algorytmów ACE i AVAS, oraz porównanie pod względem jakości predykcji modeli otrzymanych za pomocą tych metod z innymi modelami regresji.

Metody ACE i AVAS, oparte na jednoczesnej transformacji wszystkich zmiennych, są metodami nieparametrycznymi, które nie wymagają znajomości rozkładów badanych cech ani analitycznych postaci związków między nimi. Pozwalają one na wprowadzenie do modelu zmiennych zarówno metrycznych, jak i niemetrycznych. W metodach tych nie jest wymagane również definiowanie analitycznych postaci funkcji transformujących θ oraz f_i (dla $i=1, \dots, p$). Są one wyznaczone iteracyjnie w kolejnych krokach algorytmu.

2. Metoda ACE

Metoda ACE (*Alternating Conditional Expectation*) została zaproponowana w 1985 r. przez L. Breimana i J. Friedmana [1985, s. 580-619]. W metodzie tej rozważamy model addytywny postaci:

$$\theta(Y) = \sum_{i=1}^p f_i(X_i) + \varepsilon, \quad (4)$$

gdzie θ oraz f_i (dla $i=1, \dots, p$) są funkcjami mierzalnymi o skończonych wartościach momentów zwykłych drugiego rzędu i zerowej wartości oczekiwanej:

$$E\theta^2(Y) < \infty, \quad Ef_i^2(X_i) < \infty \quad \text{dla } i=1, \dots, p, \quad (5)$$

$$E\theta(Y) = 0, \quad Ef_i(X_i) = 0 \quad \text{dla } i=1, \dots, p. \quad (6)$$

Nie zakładamy monotoniczności transformacji θ oraz f_i ani tym bardziej znajomości analitycznych postaci tych funkcji.

Metoda ACE polega na jednoczesnym transformowaniu zarówno zmiennej objaśnianej, jak i zmiennych objaśniających tak, by minimalizować współczynnik zbieżności dany wzorem:

$$e^2(\theta, f_1, \dots, f_p) = \frac{E\left(\left(\theta(Y) - \sum_{i=1}^p f_i(X_i)\right)^2\right)}{E\theta^2(Y)}. \quad (7)$$

Współczynnik ten informuje, jaka część wartości przetransformowanej zmiennej $\theta(Y)$ nie została wyjaśniona przez wyrażenie $\sum_{i=1}^p f_i(X_i)$.

Oznaczmy przez $\theta^*, f_1^*, \dots, f_p^*$ rozwiązanie optymalne zadania minimalizacji współczynnika zbieżności (7), zatem:

$$(\theta^*, f_1^*, \dots, f_p^*) = \arg \min_{\theta, f_1, \dots, f_p} e^2(\theta, f_1, \dots, f_p). \quad (8)$$

W roku 1985 L. Breiman i J. Friedman udowodnili istnienie rozwiązania optymalnego zadania (8) (zob. [Breiman, Friedman 1985, s. 580-619]).

Zauważmy, że bez straty ogólności możemy założyć, że

$$E\theta^2(Y) = 1, \quad (9)$$

wtedy zadanie minimalizacji współczynnika zbieżności przybiera postać:

$$e^2(\theta, f_1, \dots, f_p) = E\left(\left(\theta(Y) - \sum_{i=1}^p f_i(X_i)\right)^2\right) \rightarrow \min. \quad (10)$$

Wyznaczenie rozwiązania zadania (10) następuje w sposób iteracyjny poprzez zastosowanie algorytmu ACE. W kolejnych krokach tego algorytmu szukamy optymalnych transformacji zarówno zmiennej objaśnianej, jak i zmiennych objaśniających, które minimalizują współczynnik zbieżności (10). W celu wyjaśnienia konstrukcji algorytmu ACE zaprezentujemy dwa jego istotne etapy.

a) Zauważmy, że jeśli znane są przekształcenia zmiennych objaśniających $f_i(X_i)$ (dla $i=1, \dots, p$), to rozwiązanie zadania minimalizacji (10) przybiera postać:

$$\hat{\theta}(Y) = \frac{E\left(\sum_{i=1}^p f_i(X_i) | Y\right)}{\left\|E\left(\sum_{i=1}^p f_i(X_i) | Y\right)\right\|}, \quad (11)$$

gdzie $\|\cdot\| = \sqrt{E(\cdot)^2}$.

b) Analogicznie, jeśli znane są transformacje:

$$\theta(Y) \text{ oraz } f_1(X_1), \dots, f_{k-1}(X_{k-1}), f_{k+1}(X_{k+1}), \dots, f_p(X_p),$$

to rozwiązanie zadania (10) można zapisać w postaci:

$$\hat{f}_k(X_k) = E\left(\theta(Y) - \sum_{i \neq k} f_i(X_i) | X_k\right). \quad (12)$$

Pełny algorytm ACE można przedstawić w następujących etapach (zob. [Breiman, Friedman 1985, s. 580-619]):

1. Przyjmujemy:

$$\theta(Y) = \frac{Y}{\|Y\|} \text{ oraz } f_i(X_i) = 0 \text{ dla } i=1, \dots, p.$$

2. Wykonujemy kolejne kroki:

a) Dla $k=1, 2, \dots, p$ wyznaczamy estymatory transformacji f_k :

$$\hat{f}_k(X_k) = E\left(\theta(Y) - \sum_{i \neq k} f_i(X_i) \mid X_k\right)$$

oraz przyjmujemy, że

$$f_k(X_k) = \hat{f}_k(X_k).$$

Powtarzamy krok 2a, dopóki maleje wartość współczynnika zbieżności $e^2(\theta, f_1, \dots, f_p)$.

b) Wyznaczamy estymator transformacji zmiennej objaśnianej:

$$\hat{\theta}(Y) = \frac{E\left(\sum_{i=1}^p f_i(X_i) \mid Y\right)}{\left\|E\left(\sum_{i=1}^p f_i(X_i) \mid Y\right)\right\|}$$

oraz przyjmujemy, że

$$\theta(Y) = \hat{\theta}(Y).$$

Dla wyznaczonych transformacji obliczamy współczynnik zbieżności $e^2(\theta, f_1, \dots, f_p)$ i powtarzamy krok 2, dopóki wartość tego współczynnika maleje.

3. Jako rozwiązanie optymalne przyjmujemy:

$$\theta^*(Y) = \theta(Y) \text{ oraz } f_i^*(X_i) = f_i(X_i) \text{ dla } i = 1, \dots, p.$$

L. Breiman i J. Friedman pokazali, że algorytm ACE jest zbieżny do rozwiązania optymalnego zadania minimalizacji (10) (zob. [Breiman, Friedman 1985, s. 580-619]).

3. Metoda AVAS

W 1988 roku w odpowiedzi na algorytm ACE Tibshirani zaproponował nieparametryczną metodę AVAS (*Additivity and Variance Stabilization*) [Tibshirani 1988, s. 394-405]. W metodzie tej również przyjmujemy, że model regresyjny ma postać przedstawioną wzorem (4), gdzie transformacje zmiennych θ oraz f_i (dla $i = 1, \dots, p$) są funkcjami mierzalnymi spełniającymi warunki:

$$E\theta^2(Y) = 1, E f_i^2(X_i) < \infty \text{ dla } i = 1, \dots, p, \quad (13)$$

$$E\theta(Y) = 0, E f_i(X_i) = 0 \text{ dla } i = 1, \dots, p. \quad (14)$$

Dodatkowo zakładamy, że funkcja θ jest silnie rosnąca.

Celem metody AVAS jest wyznaczenie takich funkcji transformujących θ oraz f_i (dla $i = 1, \dots, p$), które *stabilizują wariancję* przekształconej zmiennej objaśnianej, czyli spełniają warunki:

$$E(\theta(Y) | X_1, \dots, X_p) = \sum_{i=1}^p f_i(X_i), \quad (15)$$

$$\text{var}\left(\theta(Y) \left| \sum_{i=1}^p f_i(X_i) \right.\right) = \text{const}. \quad (16)$$

W algorytmie AVAS szukana funkcja θ wyznaczona zostaje na podstawie *transformacji asymptotycznie stabilizującej wariancję*, wyznaczonej „metodą delt” z szeregu Taylora (zob. [Tibshirani 1988, s. 394-405]). W metodzie tej rozpatrujemy rodzinę rozkładów zmiennej losowej W o średniej równej u oraz wariancji $V(u)$. Transformacja h asymptotycznie stabilizująca wariancję przybiera wtedy postać:

$$h(t) = \int_0^t \frac{1}{\sqrt{V(u)}} du. \quad (17)$$

Jeśli wyznaczone przekształcenie zmiennej objaśnianej spełnia warunek (16), to funkcja h jest funkcją tożsamościową. Jeśli jednak warunek (16) nie jest spełniony, to nowa transformacja będąca złożeniem $h(\theta(Y))$ prowadzi do ustabilizowania wariancji. Zazwyczaj zachodzi również konieczność liniowego przekształcenia funkcji $h(\theta(Y))$ tak, by spełniała ona warunki (13) i (14).

Kolejne *etapy algorytmu AVAS* można przedstawić w następujących sposób (zob. [Harrell 2001; Tibshirani 1988, s. 394-405]):

1. Przyjmujemy:

$$a) \theta(Y) = \frac{Y - EY}{\sqrt{\text{var}(Y)}}, \quad f_i(X_i) = 0 \text{ dla } i = 1, \dots, p.$$

b) Dla $k = 1, 2, \dots, p$ wyznaczamy estymatory transformacji f_k :

$$\hat{f}_k(X_k) = E\left(\theta(Y) - \sum_{i \neq k} f_i(X_i) \mid X_k\right)$$

oraz przyjmujemy, że

$$f_k(X_k) = \hat{f}_k(X_k).$$

2. Wykonujemy następujące kroki:

a) Obliczamy:

$$\hat{m}(X_1, \dots, X_p) = \sum_{i=1}^p f_i(X_i)$$

oraz

$$V(u) = \text{var}(\theta(Y) \mid \hat{m}(X_1, \dots, X_p) = u).$$

b) Wyznaczamy transformację asymptotycznie stabilizującą wariancję:

$$h(t) = \int_0^t \frac{1}{\sqrt{V(u)}} du$$

oraz przyjmujemy, że

$$\hat{\theta}(Y) = h(\theta(Y)).$$

Nowa transformacja zmiennej objaśnianej przybiera postać:

$$\theta(Y) = \frac{\hat{\theta}(Y) - E(\hat{\theta}(Y))}{\sqrt{\text{var}(\hat{\theta}(Y))}}.$$

3. Ponownie wyznaczamy estymatory transformacji zmiennych objaśniających:

$$\hat{f}_k(X_k) = E\left(\theta(Y) - \sum_{i \neq k} f_i(X_i) \mid X_k\right) \text{ dla } k = 1, 2, \dots, p$$

oraz przyjmujemy, że

$$f_k(X_k) = \hat{f}_k(X_k).$$

4. Obliczamy wartość współczynnika:

$$R^2 = 1 - E\left(\theta(Y) - \sum_{i=1}^p f_i(X_i)\right)^2.$$

Powtarzamy kroki 2-4, dopóki wartość współczynnika R^2 wzrasta.

4. Porównanie metod ACE i AVAS z innymi metodami regresji

Do analizy przedstawionych nieparametrycznych metod ACE i AVAS przeprowadzono obliczenia na zbiorach danych: *Boston*, *Friedman1*, *Friedman2* i *Friedman3*.

Dane w zbiorze *Boston* zostały zebrane w roku 1978 przez D. Harrisona i D.L. Rubinfeld (zob. [Harrison, Rubinfeld 1978]), którzy zajmowali się badaniem zależności pomiędzy jakością życia a cenami nieruchomości w okolicach Bostonu.

Celem badań autorów było określenie wpływu zanieczyszczeń powietrza na ceny nieruchomości w okolicach Bostonu. Zbiór *Boston* zawierał 506 obserwacji, a zgromadzone dane charakteryzowane były przez 13 zmiennych objaśniających (w tym jedną zmienną niemetryczną). Zmienną zależną *MEDV* jest mediana wartości domu (w tys. dolarów), natomiast wpływ zanieczyszczeń powietrza wyraża zmienna *NOX* zawierająca wielkości stężeń tlenu azotu.

Model zbudowany przez Harrisona i Rubinfelda, oparty na transformacji zarówno zmiennej objaśnianej, jak i zmiennych objaśniających, miał postać:

$$\begin{aligned} \log(MEDV) = & a_1 + a_2(RM)^2 + a_3AGE + a_4 \log(DIS) + a_5 \log(RAD) + \\ & + a_6TAX + a_7PTRATIO + a_8(BLACK - 0,63)^2 + a_9 \log(LSTAT) + \\ & + a_{10}CRIM + a_{11}ZN + a_{12}INDUS + a_{13}CHAS + a_{14}(NOX)^2. \end{aligned} \quad (18)$$

Współczynniki a_i tego modelu zostały oszacowane metodą najmniejszych kwadratów.

Dla ilustracji przedstawionych nieparametrycznych metod ACE i AVAS na zbiorze *Boston* zbudowane zostały odpowiednie modele regresji. Podobnie jak w modelu opisanym równaniem (18) zastosowano w nich transformacje zmiennych, jednak przekształcenia te nie zostały zadane *a priori*, lecz wyznaczone iteracyjnie w kolejnych krokach obu algorytmów. Wszystkie obliczenia i wykresy wykonano wykorzystując zaimplementowane w pakiecie statystycznym „R” funkcje *ace* i *avas* (zob. [Harrell 2001; Venables, Ripley 1997]).

Wykresy 1 i 2 przedstawiają funkcje transformujące dla wybranych zmiennych, wyznaczone iteracyjnie odpowiednio w algorytmach ACE i AVAS.

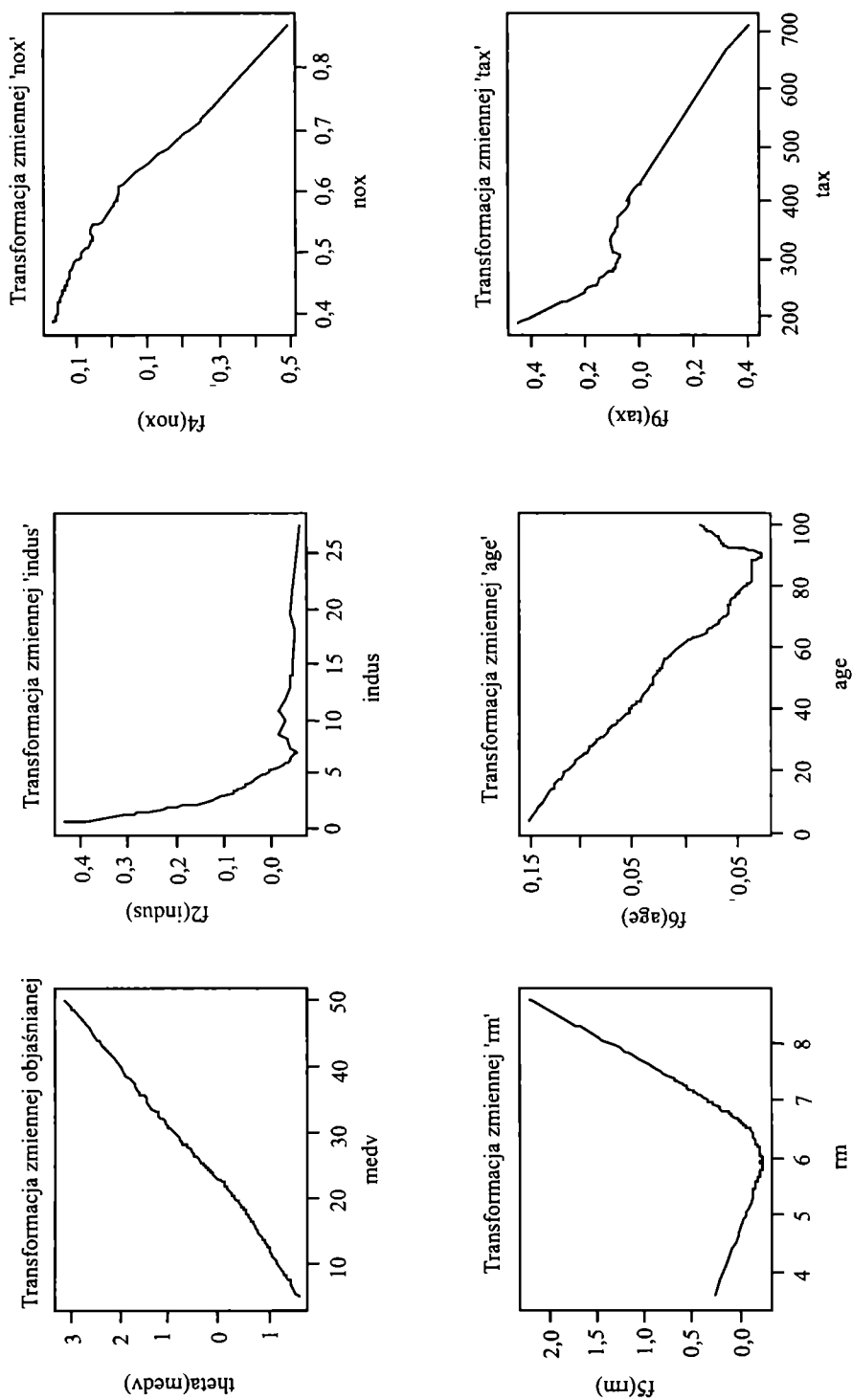
Zbiory *Friedman1*, *Friedman2* i *Friedman3* są sztucznymi zbiorami danych, wygenerowanymi komputerowo za pomocą funkcji dostępnych w pakiecie *ml-bench* programu statystycznego „R”, zgodnie z koncepcją Friedmana opisaną w pracy [Friedman 1991]. Zbiory te są standardowo wykorzystywane do porównywania własności różnych modeli regresji. Tabela 1 przedstawia wybrane charakterystyki zbiorów.

Tabela 1. Charakterystyka zbiorów danych wykorzystanych w analizie

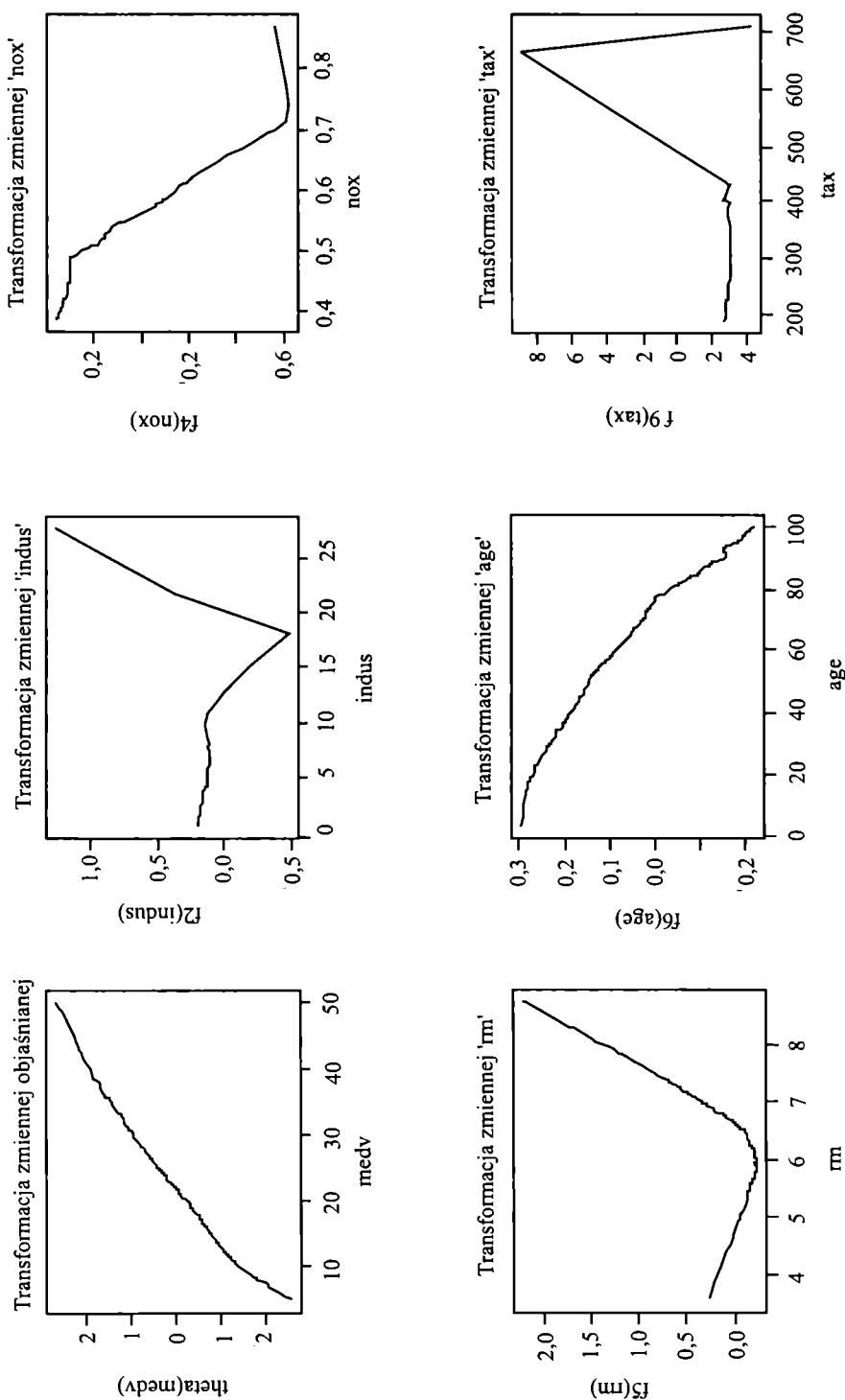
Zbiór danych	Liczebność	Liczba zmiennych	Postać funkcji
<i>Friedman 1</i>	500	10	$f(\mathbf{x}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0,5)^2 + 10x_4 + 5x_5 + e$
<i>Friedman 2</i>	500	4	$f(\mathbf{x}) = \sqrt{x_1^2 + \left(x_2 x_3 - \frac{1}{x_2 x_4}\right)^2} + e$
<i>Friedman 3</i>	500	4	$f(\mathbf{x}) = \arctg \frac{x_2 x_3 - \frac{1}{x_2 x_4}}{x_1} + e$

Źródło: opracowanie własne.

Składnik losowy e jest szumem gaussowskim o rozkładzie $N(0, s)$, gdzie s jest parametrem zadawanym przez użytkownika.



Rys. 1. Wybrane funkcje transformujące wyznaczone w algorytmie ACE dla zbioru *Boston*
 Źródło: opracowanie własne za pomocą pakietu statystycznego „R”.



Rys. 2. Wybrane funkcje transformujące wyznaczone w algorytmie AVAS dla zbioru *Boston*
 Źródło: opracowanie własne za pomocą pakietu statystycznego „R”.

Procedura badawcza obejmowała następujące etapy:

1. Dla każdego ze zbiorów wyznaczono część uczącą i testową.

a. Zbiór *Boston* został losowo podzielony na część uczącą (80% obserwacji) oraz część testową (20%).

b. Zbiory uczące i testowe dla zbiorów *Friedman1*, *Friedman2*, *Friedman3* zostały wygenerowane komputerowo niezależnie od siebie. Do zbiorów uczących, zawierających po 500 elementów dodany został szum gaussowski na poziomie 10% zmienności mierzonej wariancją. Zbiory testowe o liczebności 200 elementów nie zawierają szumu.

2. Na każdym zbiorze uczącym za pomocą algorytmów ACE i AVAS oraz odpowiadających im funkcji z pakietu „R” zbudowano odpowiednie modele regresji.

3. Dla każdego z modeli, wykorzystując zaimplementowane w pakiecie „R” funkcje: *areg.boot* i *predict*, wyznaczono teoretyczne wartości zmiennej objaśnianej, obliczone za pomocą transformacji odwrotnej θ^{-1} .

4. Dla zbudowanych modeli na zbiorze testowym obliczono wartości współczynnika determinacji R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (19)$$

5. Otrzymane modele zostały porównane pod względem jakości dopasowania z modelami wyznaczonymi za pomocą: klasycznej metody najmniejszych kwadratów oraz innych nieparametrycznych metod regresji, takich jak:

- wielowymiarowa metoda krzywych sklepanych MARS,
- metoda rzutowania PPR,
- addytywna metoda drzew regresyjnych MART.

Uzyskane wyniki przedstawia tab. 2.

Tabela 2. Wartości współczynnika R^2 wyznaczonego na zbiorze testowym dla odpowiednich modeli regresji

R^2	Regresja liniowa	MARS	PPR	MART	ACE	AVAS
Boston	0,682	0,775	0,879	0,746	0,713	0,645
Friedman1	0,739	0,982	0,839	0,905	0,881	0,883
Friedman2	0,876	0,990	0,983	0,989	0,915	0,518
Friedman3	0,568	0,880	0,925	0,898	0,870	0,866

Źródło: opracowanie własne za pomocą pakietu statystycznego „R”.

Wyniki przedstawione w tab. 2 pokazują, że największe wartości współczynnika determinacji otrzymujemy dla modeli zbudowanych metodą krzywych skleja-nych MARS lub metodą rzutowania PPR. Wiele przeprowadzonych analiz potwierdza te rezultaty, modele uzyskane za pomocą metod PPR oraz MARS dają najczęściej najmniejsze błędy prognoz. Natomiast dla modeli zbudowanych metodą ACE zazwyczaj otrzymujemy przeciętne wartości współczynnika determinacji obliczonego na zbiorze testowym. W przypadku zbiorów *Boston* i *Friedman2* wyniki R^2 są najniższe dla modeli uzyskanych metodą AVAS.

5. Podsumowanie

Metody ACE i AVAS są nieparametrycznymi metodami regresji opartymi na jednoczesnej transformacji wszystkich zmiennych. W metodach tych nie jest wymagana znajomość rozkładów wprowadzanych do modelu cech i mogą to być zmienne zarówno metryczne, jak i niemetryczne. Ograniczenie liczby założeń pozwala na dużo szersze zastosowanie tych metod niż np. klasycznej metody najmniejszych kwadratów.

Funkcje transformujące, wykorzystywane w modelu, muszą być funkcjami mierzalnymi. Nie wymagane jest podawanie analitycznej postaci tych transformacji, ponieważ są one wyznaczone iteracyjnie w kolejnych krokach algorytmów. W metodzie AVAS zakładamy jedynie, że funkcja θ przekształcająca zmienną objaśnianą jest silnie rosnąca.

Krótką analizę przedstawionych metod oraz wyniki zaprezentowane w tab. 2 pokazują, że modele zbudowane za pomocą metod ACE i AVAS wobec innych metod regresji nie dają wysokich wartości miar dopasowania. Konstrukcja algorytmów ACE i AVAS prowadzi do uzyskania modeli regresyjnych, które często są zbyt dobrze dopasowane do danych ze zbioru uczącego, co powoduje, że dają one stosunkowo duże błędy predykcji. Wadę tę można częściowo przezwyciężyć, wykorzystując do budowy modelu losowanie ze zwracaniem obiektów ze zbioru uczącego, czyli metodę *bootstrap*. Koncepcja ta została przedstawiona w pracy F. Harrella (zob. [Harrell 2001]).

Literatura

- Breiman L., Friedman J., *Estimating Optimal Transformations for Multiple Regression and Correlation (with Discussion)*, „Journal of the American Statistical Association” 1985 nr 80, s. 580-619.
 Friedman J., *Multivariate Adaptive Regression Splines*, „Annals of Statistics” 1991 nr 19, s. 1-141.
 Harrell F., *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York 2001.
 Harrison D., Rubinfeld D.L., *Hedonic Prices and the Demand for Clean Air*, „Journal of Environmental Economics and Management” 1987 nr 8.
 Hastie T., Tibshirani R., *Generalized Additive Models*, Chapman and Hall, London 1990.

Tibshirani R., *Estimating Transformations for Regression Via Additivity and Variance Stabilization*, „Journal of the American Statistical Association” 1988 nr 83, s. 394-405.
Venables W., Ripley B., *Modern Applied Statistics with S-PLUS*, Springer-Verlag, New York 1997.

VARIABLES TRANSFORMATION MODELS FOR REGRESSION

Summary

In the paper two nonparametric methods for regression, using the variables transformation, are presented. One of them is the ACE (alternating conditional expectation) algorithm of Breiman and Friedman. The other is Tibshirani's AVAS (additivity and variance stabilization) method. A numerical example for illustration and comparison is given.