PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

MODELE SEGMENTOWE W ANALIZIE REGRESJI

Streszczenie: Modele segmentowe to modele wykorzystujące rekurencyjny podział wielowymiarowej przestrzeni zmiennych na podprzestrzenie (segmenty). W każdym z tych segmentów jest budowany model lokalny, np. liniowy, a następnie modele te są łączone w jeden model globalny.

Celem artykułu jest przedstawienie propozycji nowej metody budowy modeli segmentowych, która wykorzystuje podejście wielomodelowe. W pracy pokazano także zastosowanie różnych typów modeli segmentowych w analizie regresji oraz porównanie dokładności ich dopasowania do danych.

1. Wstęp

Modele segmentowe są coraz częściej wykorzystywane w analizie regresji, głównie ze względu na ich znakomite własności, tj. elastyczność i dobre dopasowanie do danych. Jest to możliwe głównie dzięki temu, że w procesie ich budowy wykorzystywana jest metoda rekurencyjnego podziału wielowymiarowej przestrzeni zmiennych.

Klasa modeli segmentowych obejmuje drzewa regresyjne, modele krzywych sklejanych MARS oraz modele hybrydowe. W artykule zaproponowano dodatkowo nową metodę budowy tego typu modeli, wykorzystującą podejście wielomodelowe [Gatnar 2008].

Poniżej zostały omówione najważniejsze własności poszczególnych rodzajów modeli segmentowych wykorzystywanych w analizie regresji oraz wyniki porównań dokładności predykcji dla wybranych zbiorów danych. Obliczenia wykonano w środowisku programu **R**, który służy do prowadzenia zaawansowanych analiz statystycznych [Walesiak, Gatnar 2009].

2. Metoda rekurencyjnego podziału a modele segmentowe

Metoda rekurencyjnego podziału dzieli sekwencyjnie wielowymiarową przestrzeń zmiennych $\mathbf{X}^L$ na podprzestrzenie R_k (segmenty) aż do chwili, gdy zmienna zależna Y osiągnie w każdej z nich minimalny poziom zróżnicowania (mierzony za pomocą odpowiedniej funkcji straty). Metoda ta była stosowana w statystyce już przez

Morgana i Sonquista [1963]. Jej wykorzystanie w analizie dyskryminacyjnej i regresji przedstawili Breiman i in. [1984]. W języku polskim zagadnienie budowy modeli za pomocą metody rekurencyjnego podziału jest omawiane w pracy Gatnara [2001].

Algorytm metody rekurencyjnego podziału składa się z kilku etapów:

1. Sprawdź, czy wszystkie obiekty w przestrzeni zmiennych $\mathbf{X}^L$ są jednorodne ze względu na zmienną *Y*. Jeżeli tak, to zakończ pracę.

2. W przeciwnym wypadku sprawdź wszystkie możliwe podziały przestrzeni zmiennych $\mathbf{X}^L$, tj. według każdej zmiennej $X_1,...,X_L$ oraz każdego sposobu ich dyskretyzacji, na rozłączne podprzestrzenie (segmenty) R_k.

3. Dokonaj oceny każdego z tych podziałów i wybierz najlepszy z nich.

4. Podziel przestrzeń zmiennych $\mathbf{X}^L$ na podprzestrzenie zgodnie z wybranym, najlepszym podziałem i następnie wykonaj kroki 1-4 dla każdej z podprzestrzeni.

W ramach omawianej metody model jest tworzony nie globalnie, lecz przez złożenie modeli lokalnych, zbudowanych w każdym z *K* rozłącznych segmentów, na jakie dzielona jest wielowymiarowa przestrzeń zmiennych $\mathbf{X}^L$:

$$Y = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_k(\mathbf{X}) . \tag{1}$$

W szczególnym przypadku $\alpha_0 = 0$, a modele lokalne $g_k(\mathbf{X})$ mogą mieć najprostszą postać, tj. stałej, jak to ma miejsce dla drzew klasyfikacyjnych i regresyjnych:

$$g_k(\mathbf{x}_i) = I(\mathbf{x}_i \in R_k) , \tag{2}$$

gdzie R_k ($k = 1,...,K$) to podprzestrzenie (segmenty) przestrzeni $\mathbf{X}^L$, α_k – parametry modelu, *I* zaś jest funkcją wskaźnikową.

Każdy z segmentów R_k jest definiowany przez jego granice w przestrzeni $\mathbf{X}^L$, które dla zmiennych metrycznych $X_1,...,X_L$ można przedstawić jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}\right) , \tag{3}$$

gdzie wartości $v_{kl}^{(d)}$ oraz $v_{kl}^{(g)}$ oznaczają odpowiednio jego górną i dolną granicę w *l*-tym wymiarze przestrzeni zmiennych.

W przypadku, gdy zmienne $X_1,...,X_L$ mają charakter niemetryczny, podprzestrzeń R_k można zdefiniować jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(x_{il} \in B_{kl}\right) , \tag{4}$$

gdzie B_{kl} to podzbiór zbioru kategorii zmiennej X_l, tj. $B_{kl} \subseteq V_l$.

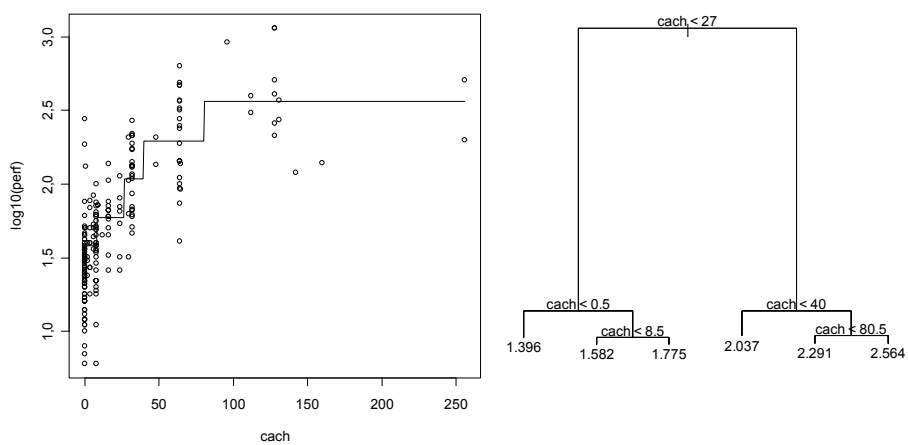
3. Drzewa regresyjne

Graficzną postacią modelu (1) z funkcjami lokalnymi $g_k(\mathbf{X})$ w postaci stałych (2) jest drzewo regresyjne. Jego parametry są wyznaczane za pomocą formuły:

$$\alpha_k = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} y_i \,, \tag{5}$$

gdzie: $N(k)$ to liczba obserwacji należących do segmentu R_k. Innymi słowy – parametr ten jest wartością przeciętną Y dla obserwacji znajdujących się w segmencie R_k.

Ilustracją może być drzewo regresyjne dla wskaźnika szybkości pracy jednostki centralnej komputera[1] (*perf*), gdzie zmienną objaśniającą była zmienna *cach* (wielkość tzw. pamięci podręcznej w kilobajtach)[2]. Jego graficzna postać znajduje się na rys. 1.



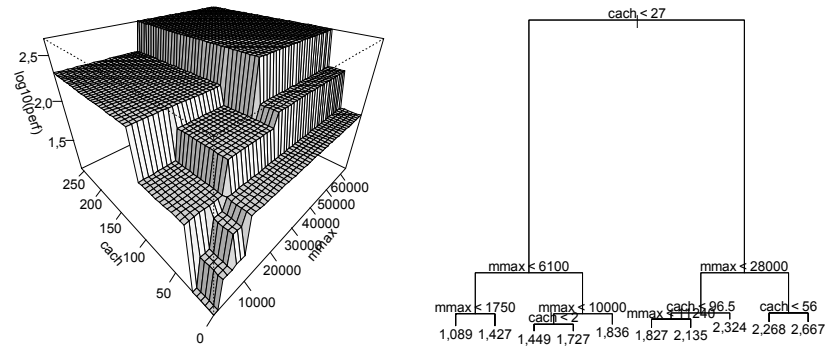
Rys. 1. Drzewo regresyjne dla jednej zmiennej objaśniającej

Źródło: opracowanie własne.

W przypadku tego samego modelu regresji, lecz z dwiema zmiennymi objaśniającymi: *cach* i *mmax* (największa pojemność pamięci operacyjnej w kilobajtach), podział przestrzeni został pokazany na rys. 2.

[1] W porównaniu z komputerem IBM 370/158-3.

[2] Dane te znajdują się w zbiorze CPUS, zawierającym dane dotyczące parametrów pracy 209 komputerów stacjonarnych, dostępnym w repozytorium baz danych Uniwersytetu Kalifornijskiego [Blake, Keogh, Merz 1988].



Rys. 2. Drzewo regresyjne dla dwóch zmiennych objaśniających

Źródło: opracowanie własne.

4. Krzywe sklejane

W modelach krzywych sklejanych (splines), np. MARS [Friedman 1991], modele lokalne $g_k(\mathbf{x})$ mają postać iloczynów tensorowych wielomianów niskich rzędów:

$$g_k(\mathbf{x}) = \prod_{j=1}^{d} b_q^*(\mathbf{x}, \mathbf{t}), \tag{6}$$

gdzie $b_q^*(\mathbf{x}, \mathbf{t})$ jest jednostronnie uciętym wielomianem stopnia q w postaci jednej z funkcji:

$$\begin{aligned} b_q^+(\mathbf{x}, \mathbf{t}) &= [x - t]_+^q \\ b_q^-(\mathbf{x}, \mathbf{t}) &= [x - t]_-^q \end{aligned}, \tag{7}$$

przy czym:

$$[x]_+ = \max\{0, x\} \text{ oraz } [x]_- = \min\{0, x\}. \tag{8}$$

W uproszczeniu funkcje składowe w segmentach mają postać:

$$g_k(\mathbf{x}) = \prod_{j=1}^{d} [u_{kj} \cdot (x_{kj} - t_{kj})]_+, \tag{9}$$

gdzie $u_{kj} = \{-1, 1\}$, d zaś oznacza liczbę interakcji między zmiennymi.

Algorytm budowy modelu segmentowego składa się z dwóch etapów: wprowadzania zmiennych do modelu oraz ich eliminacji.

- Etap wprowadzania, gdy dołączane są te funkcje sklejane, które w największym stopniu redukują błąd i zachowują postać iloczynu tensorowego.

- Etap eliminacji, gdy usuwane są te składowe, które redukują złożoność modelu, nawet kosztem niewielkiego wzrostu błędu.

Do oceny jakości modelu stosowane jest uogólnione kryterium sprawdzania krzyżowego:

$$GCV(k) = \frac{\sum_{i=1}^{N} \left(y_i - \hat{g}_k(\mathbf{x}_i)\right)^2}{N\left[1 - \frac{C(k)}{N}\right]^2}, \tag{10}$$

gdzie $C(k)$ jest pewną miarą złożoności modelu, np. liczbą liści drzewa.

Ilustracją może być model regresji dla cen nieruchomości w okolicach Bostonu[3] zbudowany metodą krzywych sklejanych za pomocą funkcji `polymars`. Zmiennymi objaśniającymi są w tym modelu wszystkie pozostałe zmienne znajdujące się w tym zbiorze.

Równanie modelu w postaci krzywych sklejanych ma postać:

$$\begin{aligned} \text{medv} = {} & 59,46 - 3,61 \cdot lstat + 1,34 \cdot [lstat - 6,12]_+ + 2,79 \cdot ptratio + \\ & + 18,34 \cdot rm + 3,16 \cdot [rm - 6,43]_+ - 3,34 \cdot crim - 0,51 \cdot rm \cdot ptratio - \\ & - 68,11 \cdot dis + 43,18 \cdot [dis - 1,55]_+ + 13,11 \cdot nox - \\ & + 0,72 \cdot [lstat\text{-}23,34]_+ + 1,61 \cdot crim \cdot lstat - 15,298 \cdot nox \cdot rm + \\ & + 0,53 \cdot [lstat - 21,24]_+ + 1,024 \cdot rad - 0,093 \cdot tax + 0,088 \cdot [tax\text{ - }233]_+ - \\ & - 0,124 \cdot rad \cdot lstat + 0,114 \cdot rad \cdot [lstat - 6,29]_+ . \end{aligned} \tag{11}$$

5. Modele hybrydowe

Model lokalny $g_k(\mathbf{X})$ może być także modelem klasy GLM, a szczególnie prostym modelem liniowym:

$$g_k(\mathbf{X}) = \beta_0 + \sum_{l=1}^{L} \beta_l X_l \,. \tag{12}$$

[3] Dane te znajdują się w zbiorze BOSTON [Harrison, Rubinfeld 1978], który jest również benchmarkowym zbiorem danych, dostępnym w repozytorium Uniwersytetu Kalifornijskiego [Blake, Keogh, Merz 1988]. Zawiera on informacje o wpływie różnych czynników na ceny 506 nieruchomości na przedmieściach Bostonu. Zmienną zależną jest mediana wartości domu w tys. dolarów (*medv*), zmiennymi objaśniającymi zaś są: koncentracja tlenku azotu (*nox*), wskaźnik przestępstw (*crim*), wielkość podatku od nieruchomości (*tax*), procent ludności murzyńskiej (*b*), procent budynków zbudowanych przed 1940 r. (*age*), dostęp do autostrady (*rad*), wskaźnik uczniów przypadających na jednego nauczyciela (*ptratio*), ważona odległość od pięciu centrów zatrudnienia w Bostonie (*dis*), dostęp do rzeki Charles River (*chas*), odsetek terenów nieprzeznaczonych do celów handlowych (*indus*), odsetek terenów zamieszkałych (*zn*), procent populacji o niskim statusie społecznym (*lstat*), średnia liczba pokoi w domu (*rm*).

Wtedy tego typu modele nazywane są modelami hybrydowymi, ponieważ łączą podejście nieparametryczne (drzewa) z parametrycznym. Czasami stosowane jest także określenie „drzewa funkcyjne" [Gama 2004].

W tym rozwiązaniu, w odróżnieniu od klasycznego, najpierw wyznaczana jest zmienna, która definiuje podział, a następnie dokonywana jest jej dyskretyzacja. Algorytm rekurencyjnego podziału ma wtedy postać:

1. Zbuduj model lokalny (liniowy) w segmencie R_k.

2. Oceń stabilność parametrów tego modelu. Jeżeli niestabilność występuje, wybierz zmienną X_l, przy której stoi najbardziej niestabilny parametr. W przeciwnym wypadku zakończ pracę.

3. Dokonaj dyskretyzacji zmiennej X_l.

4. Dokonaj podziału przestrzeni na segmenty według zmiennej X_l i przejdź do kroku 1.

Zeileis, Hothorn i Hornik [2007] zastosowali do budowy drzewa kryterium stabilności parametrów modeli lokalnych, które są estymowane metodą najmniejszych kwadratów lub metodą największej wiarygodności. Do testowania stabilności parametrów w kroku 2 stosowany jest test fluktuacji zaproponowany przez Zeileisa, Hothorna i Hornika [2007]. Wykorzystuje on statystykę:

$$W_l(t) = \frac{1}{\sqrt{N \cdot \hat{\boldsymbol{\Sigma}}}} \cdot \sum_{i=1}^{N \cdot t} \phi(X_{il}), \tag{13}$$

która jest zbieżna z procesem Browna w przypadku prawdziwości hipotezy zerowej o stabilności parametrów modelu. We wzorze (13) $\hat{\boldsymbol{\Sigma}}$ jest estymatorem macierzy wariancji i kowariancji, φ zaś jest funkcją oceny, dającą rangę obserwacji X_{il}.

Oczywiście można także zastosować inne, podobne testy stabilności parametrów, znane w ekonometrii CUSUM [Ploberger, Krämer 1992] lub test Nybloma [1989] itp.

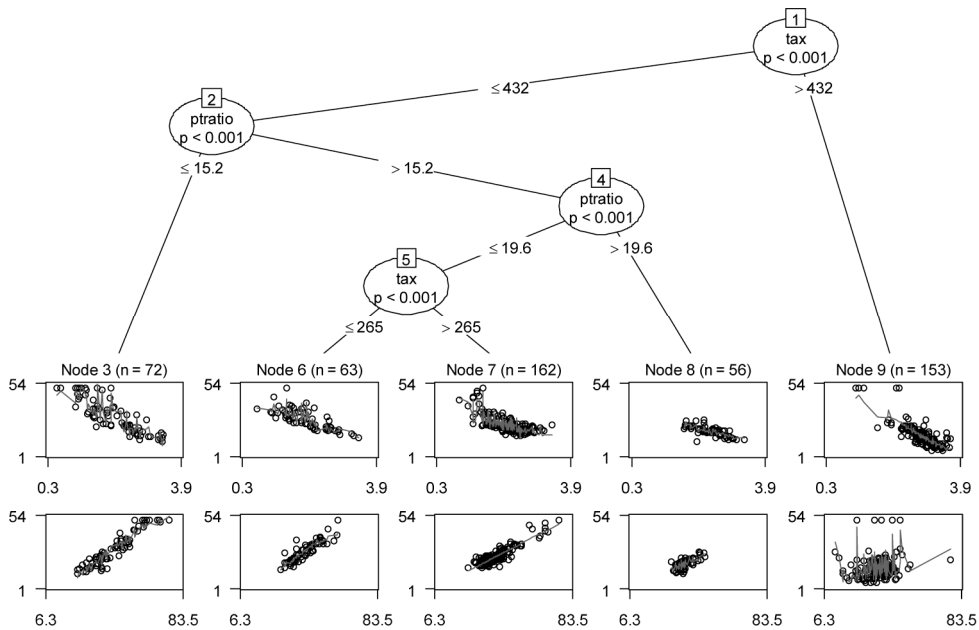
Hybrydowy model regresji dla zbioru BOSTON, w którym modele wewnętrzne (liniowe) tworzą zmienne: *lstat* oraz *rm*, został pokazany na rys. 3. Wykorzystano w tym celu polecenie `mob` z pakietu `party` w programie **R**.

Uzyskany model należy interpretować w ten sposób, że przestrzeń zmiennych została podzielona na 5 segmentów za pomocą dwóch zmiennych: wielkości podatku od nieruchomości (*tax*) oraz wskaźnika uczniów przypadających na jednego nauczyciela (*ptratio*). W każdym z nich zbudowano model liniowy, np. w segmencie 3 model lokalny ma postać:

```
1) tax <= 432; criterion = 1, statistic = 115.364
 2) ptratio <= 15.2; criterion = 1, statistic = 50.482
 3)* weights = 72
Terminal node model
Linear model with coefficients:
(Intercept) lstat rm
 9.2349 -4.9391 0.6859,
```

co oznacza funkcję:

$$g_3(\mathbf{X}) = 9{,}2349 - 4{,}9391 \cdot lstat + 0{,}6859 \cdot rm\,. \tag{14}$$



Rys. 3. Hybrydowy model regresji

Źródło: opracowanie własne.

A zatem wzrost liczby mieszkańców o niskim statusie społecznym w sąsiedztwie (*lstat*) powoduje spadek ceny nieruchomości, a wzrost liczby pokoi (*rm*) powoduje wzrost ceny nieruchomości. Wyjątkiem jest jedynie segment 9 (największe nieruchomości), w którym parametr przy zmiennej *rm* także jest ujemny (co prawda jest on jedynie nieznacznie większy od zera).

6. Propozycja nowej metody

Nowa metoda budowy modeli segmentowych wykorzystuje podejście wielomodelowe, które polega na łączeniu modeli zbudowanych na podstawie różnych zbiorów uczących [Gatnar 2008]. Model te są budowane właśnie w różnych segmentach wielowymiarowej przestrzeni zmiennych, uzyskanych za pomocą metody rekurencyjnego podziału. Algorytm jest prosty i dosyć intuicyjny:

1. Podziel przestrzeń zmiennych na M podprzestrzeni (segmentów), wybierając zmienne definiujące podział w sposób losowy.

2. Dokonaj dyskretyzacji każdej zmiennej, stosując proste kryterium liczebności segmentu (np. nie mniej niż 20% obserwacji).

3. Utwórz w każdym z segmentów model w postaci drzewa regresyjnego (D_m) o niewielkiej liczbie węzłów, np. $K = 4$, co realizuje polecenie `tree(best=4)`.

4. Dokonaj agregacji modeli lokalnych za pomocą średniej ważonej.

W rezultacie powstaje podwójne drzewo (*double tree* – DT), tj. drzewo reprezentujące podział przestrzeni na segmenty, w których znajdują się proste drzewa regresyjne. Agregacja wyników predykcji tych modeli następuje przez uśrednianie z wagami (w_m), którymi jest np. liczba obiektów w poszczególnych segmentach:

$$\hat{D}^*(\mathbf{x}_i) = \sum_{m=1}^{M} w_m \hat{D}_m(\mathbf{x}_i). \tag{15}$$

7. Eksperyment

Analiza porównawcza obejmuje ocenę dopasowania czterech modeli segmentowych zbudowanych różnymi metodami do zbioru danych (10-częściowe sprawdzanie krzyżowe) za pomocą błędu średniokwadratowego (MSE). Wykorzystano w eksperymentach 10 benchmarkowych zbiorów danych stosowanych w analizie regresji, które znajdują się w repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine [Blake, Keogh, Merz 1988].

Uzyskane wyniki przedstawiono w tab. 1.

Tabela 1. Błędy średniokwadratowe dla zbiorów danych

Zbiór danych	CART	MARS	MOB	DT
Boston	254,25	231,23	225,22	**210,87**
Longley	23,23	19,23	18,98	**16,25**
Anscombe	118,76	119,23	**118,24**	118,65
Cars93	42,45	42,23	32,34	**30,56**
Traffic	278,12	257,43	223,46	**211,56**
Michelson	4,82	5,09	**4,43**	5,32
Ships	189,23	150,43	148,72	**115,38**
Whitesite	11,34	9,61	8,94	**8,32**
Cpus	89,29	76,34	**73,81**	78,87
Precip	387,43	364,11	338,62	**324,39**

Źródło: obliczenia własne za pomocą programu **R.**

Dokonano także porównania szybkości działania analizowanych procedur (polecenie `system.time()`) dla różnych zbiorów danych, uzyskując wyniki, które znajdują się w tab. 2.

Tabela 2. Czas pracy procesora

Zbiór danych	CART	MARS	MOB	DT
Boston	1,34	2,54	2,07	**1,22**
Longley	2,71	3,62	2,12	**0,88**
Anscombe	1,12	2,86	2,03	**0,76**
Cars93	0,78	1,54	1,39	**0,61**
Traffic	1,57	2,89	2,35	**1,07**
Michelson	0,56	0,73	0,84	**0,46**
Ships	**0,32**	0,68	0,62	0,46
Whitesite	**0,42**	0,98	0,82	0,51
Cpus	**0,17**	0,69	0,38	0,25
Precip	1,26	2,83	2,05	**1,06**

Źródło: obliczenia własne za pomocą programu **R.**

Wartości najmniejsze, wskazujące na optymalne rozwiązanie, są wyróżnione w tab. 1 i 2 pogrubioną czcionką.

8. Podsumowanie

Zaproponowana metoda budowy regresyjnych modeli segmentowych (DT), wykorzystująca podejście wielomodelowe, ma następujące własności:

- stosuje modele nieparametryczne,
- oddziela poszukiwanie zmiennej definiującej podział od procesu jej dyskretyzacji,
- daje na ogół lepsze dopasowanie modelu globalnego do danych,
- jest szybsza od pozostałych metod.

Niestety, uzyskany model globalny nie ma interpretacji podobnej do modelu klasycznego lub modelu w postaci drzewa regresyjnego.

Literatura

Blake C., Keogh E., Merz C.J., *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine 1988.

Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, CRC Press, London 1984.

Friedman J.H., *Multivariate adaptive regression splines*, „Annals of Statistics" 1991 vol. 19.

Gama J., *Functional Trees*, „Machine Learning" 2004 no 55, s. 219-250.

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2001.

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Harrison D., Rubinfeld D.L., *Hedonic prices and the demand for clean air*, „Journal of Environmental Economics and Management" 1978 no 5, s. 81-102.

Morgan J.N., Sonquist J.A., *Problems in the analysis of survey data: a proposal*, „Journal of the American Statistical Association" 1963 no 58, s. 417-434.

Nyblom J., *Testing for the constancy of parameters over time*, „Journal of the American Statistical Association" 1989 no 84, s. 223-230.

Ploberger W., Krämer W., *The CUSUM test with OLS residuals*, „Econometrica" 1992 no 60, s. 271-285.

Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

Zeileis A., *A unified approach to structural change tests based on ml scores, F statistics, and OLS residuals*, „Econometric Reviews" 2005 no 24, s. 445-466.

Zeileis A., Hothorn T., Hornik K., *Model-based recursive partitioning*, „Journal of Computational and Graphical Statistics" 2007 no 17(2), s. 492-514.

SEGMENTED MODELS IN REGRESSION

Summary: Segmented models are based on the recursive partitioning of multidimensional feature space into subspaces (regions). Then, in each segment a local model is built (e.g. linear model) and finally all the local models are combined into the global model.

The aim of the paper is to present a new method for building segmented models, that uses the multiple-model approach. We also discuss the results of application of different segmented models in regression and compare the goodness of fit of the models to the training data.