Joanna Trzęsiok

METODA KRZYWYCH SKŁADANYCH W BUDOWIE MODELU REGRESYJNEGO

Wprowadzenie

W niniejszym opracowaniu przedstawiono zagadnienie wyznaczania funkcji regresji z wykorzystaniem metody krzywych składanych. W metodzie tej dzielimy dziedzinę zmiennej *X* na rozłączne przedziały i dopasowujemy wielomian *n*-tego stopnia do każdego z przedziałów z osobna, dodając warunki gładkości funkcji w punktach podziału dziedziny. Opisano dwa szczególne przypadki: naturalne funkcje składane trzeciego rzędu oraz gładkie funkcje składane. Metody te zilustrowano na przykładzie rynku wydawniczego. W analizowanym przykładzie dokonano również porównania jakości modeli otrzymywanych za pomocą metody krzywych składanych oraz metody najmniejszych kwadratów.

Przedmiotem analizy regresji w przestrzeni $\mathbf{R}^2$ jest zbiór obserwacji:

$$U = \{(x_j, y_j): \ j = 1,2,\ldots,N\}$$

gdzie x_j jest realizacją zmiennej objaśniającej *X*, natomiast y_j – realizacją zmiennej zależnej *Y* (dla $j = 1,2,\ldots,N$). Celem analizy regresji jest znalezienie takiej funkcji f, która przedstawiałaby zależność zmiennej zależnej *Y* od zmiennej objaśniającej *X*:

$$Y = f(X) + \varepsilon$$

gdzie ε jest czynnikiem losowym.

Będziemy zakładać, że zależność pomiędzy zmiennymi X i Y można opisać za pomocą funkcji addytywnej. Funkcję f możemy przedstawić w postaci:

$$Y = f(X) = \alpha_0 + \sum_{i=1}^{K} \alpha_i f_i(X) + \varepsilon \tag{1}$$

Poszukujemy zatem funkcji składowych f_i dla $i = 1,2,\ldots,K$.

Rozwiązanie tego problemu, oparte na **metodzie krzywych składanych** (*splines*), polega na podziale dziedziny zmiennej X na rozłączne przedziały i dopasowywaniu funkcji do każdego z tych przedziałów z osobna.

1. Metoda krzywych składanych

W metodzie krzywych składanych dziedzinę zmiennej X dzielimy na K rozłącznych przedziałów za pomocą uporządkowanego zbioru punktów nazywanych **węzłami**:

$$\{\xi_i\}_{i=1,\ldots,K-1}$$

W każdym przedziale $\langle \xi_i, \xi_{i+1} \rangle$ za pomocą klasycznych metod regresji szukamy funkcji f_i, która jest wielomianem co najwyżej n-tego stopnia.

W najprostszej sytuacji funkcje f_i mogą być funkcjami stałymi:

$$f_i(X) = \overline{Y_i} \quad \text{dla} \quad i = 1,\ldots,K$$

gdzie $\overline{Y_i}$ jest średnią z wartości zmiennej Y należącej do i-tego przedziału. Wtedy funkcja f zazwyczaj nie jest ciągła. Jeśli jednak podwyższymy stopień wielomianów f_i oraz nałożymy na nie warunki ciągłości funkcji w węzłach:

$$\bigwedge_{i=1,..,K-1} f_i(\xi_i) = f_{i+1}(\xi_i) \tag{2}$$

oraz np. warunki ciągłości pochodnej rzędu pierwszego:

$$\bigwedge_{i=1,..,K-1} f_i'(\xi_i) = f_{i+1}'(\xi_i) \tag{3}$$

to uzyskamy funkcję o odpowiednim stopniu gładkości. Jeżeli chcemy uzyskać wyższy stopień gładkości funkcji f, to musimy analogicznie nałożyć warunki ciągłości na pochodne wyższych rzędów funkcji f.

1.1. Funkcje składane rzędu *M*

Funkcją składaną rzędu *M* nazywamy funkcję *f* złożoną z wielomianów stopnia *M,* która ma ciągłą pochodną rzędu $(M-1)$.

Można pokazać, że funkcję składaną rzędu *M* z węzłami $\{\xi_i\}_{i=1,..,K-1}$ da się przedstawić jako kombinację liniową następujących funkcji:

$$h_k(X) = X^k \qquad \text{dla } k = 0,...,M \tag{4}$$

$$h_{M+i}(X) = (X-\xi_i)_+^M \quad \text{dla } i = 1,...,K-1$$

gdzie:

$$t_+ = \begin{cases} t & \text{dla } t \geq 0 \\ 0 & \text{dla } t < 0 \end{cases}$$

Funkcje (1) są nazywane **funkcjami bazowymi**. Natomiast liczbę funkcji bazowych, oznaczaną przez df, będziemy nazywać **stopniami swobody**.

Jak widać z powyższego przedstawienia funkcji (4), liczba stopni swobody dla funkcji składanej rzędu *M* wynosi:

$$df = M + K \tag{5}$$

Najczęściej wykorzystywanymi w praktyce funkcjami składanymi są **funkcje składane trzeciego rzędu** (*cubic splines*). Są to funkcje złożone z wielomianów trzeciego stopnia spełniających warunki:

$$\bigwedge_{i=1,..,K-1} f_i(\xi_i) = f_{i+1}(\xi_i)$$

$$\bigwedge_{i=1,..,K-1} f_i'(\xi_i) = f_{i+1}'(\xi_i) \tag{6}$$

$$\bigwedge_{i=1,..,K-1} f_i''(\xi_i) = f_{i+1}''(\xi_i)$$

Funkcję składaną trzeciego rzędu z $K-1$ węzłami $\{\xi_i\}_{i=1,...,K-1}$ można przedstawić za pomocą funkcji bazowych (1) jako:

$$f(X) = \alpha_0 + \sum_{k=1}^{3} \alpha_k X^k + \sum_{i=1}^{K-1} \beta_i (X-\xi_i)_+^3 \tag{7}$$

Aby zbudować model regresyjny metodą krzywych składanych, należy wybrać stopień wielomianów składowych oraz liczbę i położenie węzłów. Jak już wspomniano, najczęściej są wykorzystywane funkcje składane trzeciego rzędu. Korzysta się także z funkcji składanych kawałkami stałych (czyli rzędu $M = 0$) lub funkcji składanych rzędu pierwszego.

Trudniejszym do rozwiązania problemem jest wybór liczby węzłów i ich położenia. Proste podejście, wykorzystywane w pakiecie statystycznym R, polega na podaniu stopni swobody df oraz stopnia funkcji bazowych *M*. Wtedy $(df - M - 1)$ węzłów zostaje wybranych jako odpowiednie kwantyle rozkładu zmiennej *X*.

1.2. Naturalne funkcje składane trzeciego rzędu

Szczególnym rodzajem funkcji składanych trzeciego rzędu są **naturalne funkcje składane** (*natural cubic splines*). Powstają one przez nałożenie na funkcję f warunków liniowości w przedziałach $\langle -\infty, \xi_1 \rangle$ oraz $\langle \xi_{K-1}, \infty)$, w których występują często wartości nietypowe, powodujące duże odchylenia wartości funkcji f. Funkcje składane trzeciego rzędu dla dużych i nietypowych wartości zmiennej *X* mogą zbyt szybko zmierzać do nieskończoności, dlatego wprowadzamy warunki liniowości w pierwszym i ostatnim przedziale dziedziny.

Powiedzieliśmy, że funkcje składane trzeciego rzędu z *K*–1 węzłami można przedstawić w postaci kombinacji liniowych funkcji bazowych:

$$f(X) = \beta_0 + \sum_{j=1}^{3} \beta_j X^j + \sum_{k=1}^{K-1} \theta_k (X - \xi_k)_+^3 \tag{8}$$

Wprowadzając warunki liniowości dla funkcji f w przedziałach $\langle -\infty, \xi_1 \rangle$ i $\langle \xi_{K-1}, \infty)$ można pokazać, że:

$$\beta_2 = 0, \quad \beta_3 = 0, \quad \sum_{k=1}^{K-1} \theta_k = 0 \quad \text{oraz} \quad \sum_{k=1}^{K-1} \xi_k \theta_k = 0 \tag{9}$$

Oznacza to, że naturalne funkcje składane z $K - 1$ węzłami można przedstawić jako kombinację $K + 1$ funkcji bazowych i tyle też wynosi liczba stopni swobody w tym modelu. Podejście to jest zaimplementowane w pakiecie *R* przez funkcję `ns`.

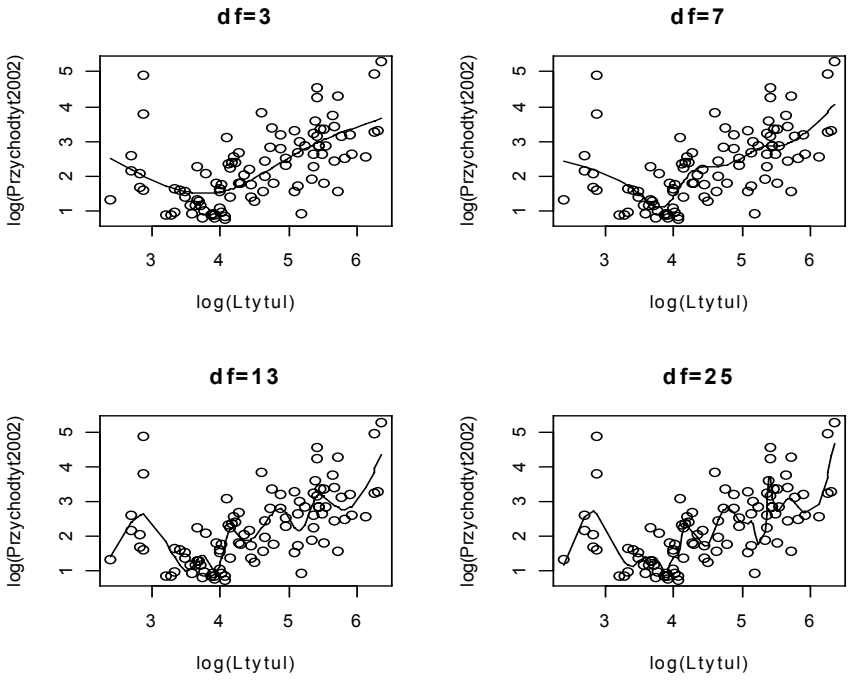
1.3. Przykład zastosowania naturalnych funkcji składanych

Zastosowanie funkcji składanych pokażemy na przykładzie rynku wydawniczego. Dane wykorzystane do skonstruowania przykładu pochodzą z rankingu wydawców przeprowadzonego przez „Rzeczpospolitą" i opublikowanego 15 maja 2003 r. w numerze 12.

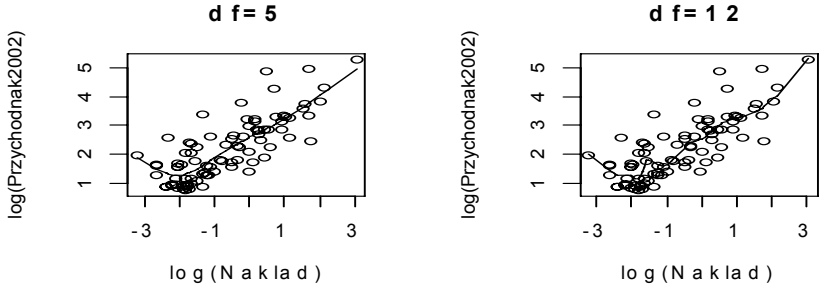
Rolę zmiennej objaśnianej będzie pełnił *Przychod2002*, czyli przychód ze sprzedaży książek w 2002 r. (w mln zł). Jako zmienne objaśniające przyjmiemy kolejno: liczbę wydawanych tytułów – *Ltytul*, łączny nakład (w mln egz.) – *Naklad* oraz liczbę zatrudnionych (w etatach) – *Lzatrud*.

Aby zastosować funkcję `ns`, musimy wybrać liczbę stopni swobody lub położenie węzłów. Wybór stopni swobody jest często dokonywany na podstawie wykresu. Dla naszego przykładu zostały przygotowane odpowiednie wykresy, wykorzystujące metodę krzywych składanych dla różnych stopni swobody. W przypadku pierwszym, opisującym zależność przychodu od liczby wydawanych tytułów (rys. 1), wydaje się, że najlepszą wartością parametru df jest 7.

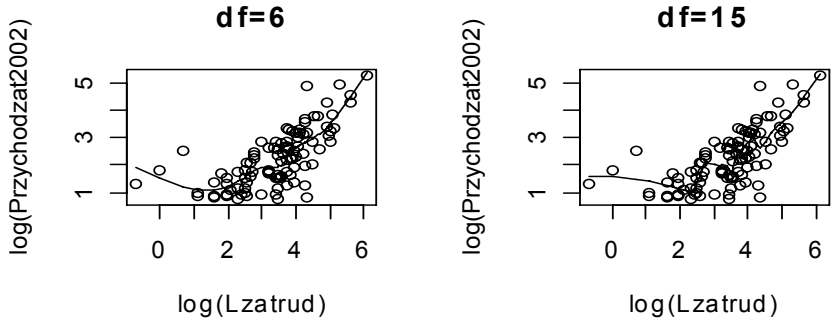
Funkcja `ns` zwraca macierz o wymiarze $N \times df$, która zawiera współczynniki funkcji bazowych dla każdej realizacji zmiennej *Ltytul*. Tak więc, np. funkcja $ns(x, df = 7)$ przy ustalonych 7 stopniach swobody tworzy model funkcji składanych złożony z 7 funkcji bazowych. Ponadto podaje rząd wielomianów składowych i położenie węzłów. W naszym przypadku funkcja składana jest więc rzędu trzeciego z 5 węzłami, które są kwantylami rozkładu zmiennej *Ltytul*.



Rys. 1. Wykres funkcji regresji, obrazującej zależność przychodu od liczby wydawanych tytułów, uzyskanej za pomocą naturalnych funkcji składanych trzeciego rzędu dla różnych stopni swobody *df*



Rys. 2. Wykres funkcji regresji, obrazującej zależność przychodu od wielkości nakładu, uzyskanej za pomocą naturalnych funkcji składanych trzeciego rzędu dla różnych stopni swobody *df*



Rys. 3. Wykres funkcji regresji, obrazującej zależność przychodu od wielkości zatrudnienia, uzyskanej za pomocą naturalnych funkcji składanych trzeciego rzędu dla różnych stopni swobody *df*

1.4. Gładkie funkcje składane

Podklasę funkcji składanych stanowią **gładkie funkcje składane** (*smoothing splines*), które unikają problemu wyboru liczby i lokalizacji węzłów. Węzłami gładkich funkcji składanych są wszystkie różne realizacje zmiennej *X*.

Niech (x_j, y_j) dla $j = 1,2,\ldots,N$ będą odpowiednio realizacjami zmiennych *X* i *Y*. Poszukiwana funkcja regresji ma być jak najlepiej dopasowana do tych danych. Jednak często wymagamy, aby funkcja regresji miała również odpowiedni stopień gładkości, np. ciągłą pochodną rzędu drugiego. Szukamy więc funkcji, która minimalizuje funkcjonał:

$$RSS(f,\lambda) = \sum_{j=1}^{N} (y_j - f(x_j))^2 + \lambda \int (f''(t))^2 \, dt \qquad (10)$$

gdzie $\lambda \in (0,\infty)$ jest parametrem gładkości, zaś f – funkcją o ciągłej pochodnej rzędu drugiego.

Pierwszy człon funkcjonału (10) odpowiada za dopasowanie funkcji f do danych, natomiast drugi jest miarą gładkości funkcji regresji. Parametr λ ustala proporcje pomiędzy dopasowaniem funkcji f do danych a jej gładkością.

Wykazano, że jedynym rozwiązaniem minimalizacji w przestrzeni funkcji z ciągłą pochodną rzędu drugiego jest naturalna funkcja składana trzeciego rzędu z węzłami w punktach x_j. Rozwiązanie to możemy zapisać jako:

$$f(X) = \sum_{i=1}^{N} N_i(X)\theta_i \tag{11}$$

gdzie N_i dla $i = 1,2,\ldots,N$ są funkcjami bazowymi dla naturalnych funkcji składanych.

Niech $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ będzie wektorem realizacji zmiennej *X*, zaś $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ wektorem realizacji zmiennej *Y*. Oznaczmy przez **N** macierz funkcji bazowych N_i, której elementy są zdefiniowane jako:

$$\mathbf{N}[j,i] = N_i(x_j) \qquad \text{dla } j = 1,2,\ldots,N, \quad i = 1,2,\ldots N$$

Łatwo pokazać, że:

$$\mathbf{y} = \mathbf{N} \cdot \boldsymbol{\theta} \tag{12}$$

gdzie $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}$ jest wektorem parametrów.

Stąd pierwszy człon funkcjonału (10) zapiszemy w postaci macierzowej jako:

$$\sum_{j=1}^{N} (y_j - f(x_j))^2 = (\mathbf{y} - \mathbf{N} \cdot \boldsymbol{\theta})^{\mathrm{T}} \cdot (\mathbf{y} - \mathbf{N} \cdot \boldsymbol{\theta}) \tag{13}$$

Natomiast jeśli przez $\boldsymbol{\Omega}_N$ oznaczymy macierz o następujących elementach:

$$\boldsymbol{\Omega}_N[j,k] = \int N_j''(t) N_k''(t)\,dt \quad \text{dla} \quad j = 1,\ldots,N, \quad k = 1,\ldots,N \tag{14}$$

to drugi człon funkcjonału (10) możemy przedstawić w postaci:

$$\int (f''(t))^2\,dt = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\Omega}_N \boldsymbol{\theta} \tag{15}$$

Zatem, zamiast szukać funkcji minimalizującej funkcjonał (10), możemy szukać wektora $\boldsymbol{\theta}$, w którym będzie osiągane minimum funkcjonału:

$$RSS(\boldsymbol{\theta}, \lambda) = (\mathbf{y} - \mathbf{N}\boldsymbol{\theta})^{\mathrm{T}} (\mathbf{y} - \mathbf{N}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\Omega}_N \boldsymbol{\theta} \tag{16}$$

Funkcjonał (16) osiąga minimum dla wektora:

$$\hat{\boldsymbol{\theta}} = (\mathbf{N}^{\mathrm{T}}\mathbf{N} + \lambda \boldsymbol{\Omega}_N)^{-1} \mathbf{N}^{\mathrm{T}} \mathbf{y} \tag{17}$$

Zaś wektor $\hat{\mathbf{f}} = \begin{bmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_N) \end{bmatrix}$, wartości funkcji $\hat{f}$ dla odpowiednich x_j, możemy zapisać jako:

$$\hat{\mathbf{f}} = \mathbf{N}\hat{\boldsymbol{\theta}} = \mathbf{N}(\mathbf{N}^{\mathrm{T}}\mathbf{N} + \lambda \boldsymbol{\Omega}_N)^{-1} \mathbf{N}^{\mathrm{T}} \mathbf{y} \tag{18}$$

Przez $\mathbf{S}_\lambda$ oznaczamy macierz występującą we wzorze (18):

$$\mathbf{S}_\lambda = (\mathbf{N}^{\mathrm{T}}\mathbf{N} + \lambda \boldsymbol{\Omega}_N)^{-1} \mathbf{N}^{\mathrm{T}} \tag{19}$$

która jest nazywana **macierzą wygładzającą** (*smoother matrix*).

Gładkie funkcje składane o N węzłach $\{x_j\}_{j=1,\ldots,N}$ mają, jako naturalne funkcje składane, *N* stopni swobody. Jednak w przypadku tych funkcji definiujemy również **efektywne stopnie swobody** df_λ jako ślad macierzy wygładzającej $\mathbf{S}_\lambda$:

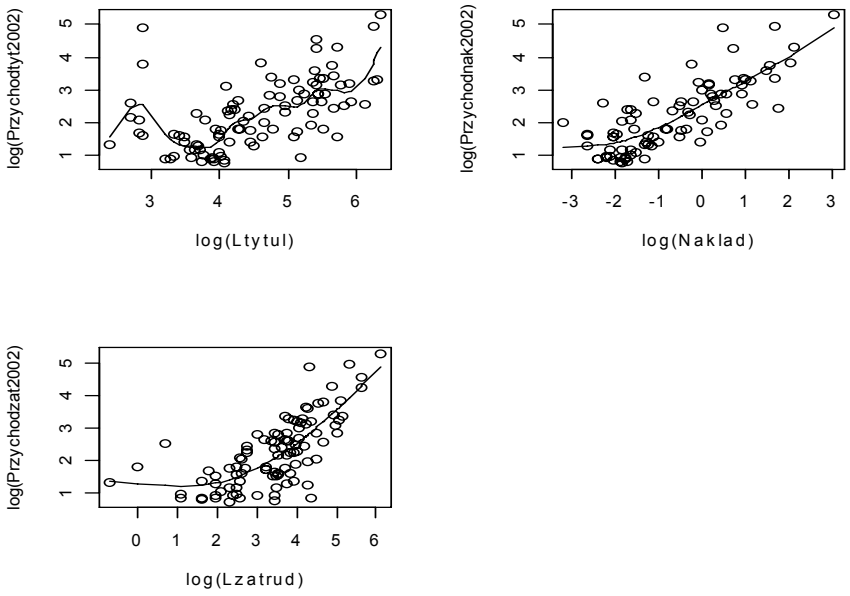
$$df_\lambda = \mathrm{tr}(\mathbf{S}_\lambda) \tag{20}$$

Przy ustalaniu stopni swobody df_λ możemy, za pomocą metod numerycznych, wyznaczyć parametr wygładzający λ.

Gładkie funkcje składane są zaimplementowane w pakiecie *R* w postaci funkcji `smooth.spline(x,y)`. Dla funkcji tej możemy opcjonalnie podać liczbę stopni swobody df_λ lub ustalić parametr gładkości λ.

1.5. Przykład zastosowania gładkich funkcji składanych

Zastosowanie gładkich funkcji składanych zilustrujemy, tak jak wcześniej, na przykładzie rynku wydawniczego. Pokażemy kształt zależności zmiennej objaśnianej *Przychod2002* kolejno od: liczby wydawanych tytułów – *Ltytul*, wielkości nakładu – *Naklad* oraz wielkości zatrudnienia – *Lzatrud*.



Rys. 4. Wykres funkcji regresji uzyskanej za pomocą gładkich funkcji składanych

Parametry modelu regresyjnego, uzyskanego metodą gładkich funkcji składanych, otrzymujemy przez wywołanie procedury:

```
smooth.spline(x =log(Ltytul),y =log(Przychod2002))
```

która w tym przypadku opisuje zależność przychodu od liczby wydawanych tytułów.

Funkcja `smooth.spline` zwraca między innymi wartość wyliczonego parametru λ oraz liczbę efektywnych stopni swobody df_λ. Natomiast parametr GCV informuje o wartości uogólnionego kryterium sprawdzania krzyżowego, za pomocą którego jest określona jakość dopasowania modelu do danych.

Dla przykładu wyznaczono dopasowanie funkcji regresji (uzyskanej metodą gładkich funkcji składanych, opisującej zależność przychodu w 2002 r.) do liczby wydawanych tytułów. Obliczono, że współczynnik R^2 dla funkcji regresji wynosi w przybliżeniu 0,55. W tab. 1 podano również wartości tego współczynnika dla funkcji regresji: liniowej, wielomianowej i wykładniczej.

Tabela 1

Dopasowanie różnych funkcji regresji wyznaczone za pomocą współczynnika R^2

Postać funkcji regresji	R^2
Funkcja liniowa	0,3038
Wielomian stopnia 3.	0,4254
Funkcja wykładnicza	0,3095
Gładka funkcja składana	0,5476

Z tab. 1 wynika, że najlepsze dopasowanie daje model oparty na gładkich funkcjach składanych i jedynie dla tego modelu współczynnik R^2 jest większy niż 0,5.

Podsumowanie

Funkcja regresji otrzymana metodą krzywych składanych jest lepiej dopasowana do danych empirycznych niż np. funkcja regresji otrzymana za pomocą klasycznej metody najmniejszych kwadratów. Atutem modeli regresyjnych opartych na krzywych składanych jest ich lokalny charakter. Metoda ta dopasowuje wielomiany co najwyżej *n*-tego stopnia w przedziałach dziedziny zmiennej *X* pomiędzy kolejnymi węzłami.

Wadą tej metody jest to, że otrzymamy duże błędy predykcji dla argumentów spoza zakresu zmienności zmiennej *X,* ponieważ np. naturalne funkcje składane przybliżają wartości teoretyczne dla argumentów *x*, mniejszych od pierwszego węzła lub większych od ostatniego, za pomocą funkcji liniowej.

Metoda krzywych składanych jest jednak bardzo dobrym narzędziem, jeśli chodzi o predykcję wewnątrz przedziału zmienności realizacji zmiennej objaśniającej. Tym samym możemy tę metodę wykorzystywać np. do uzupełniania braków danych, co stanowi istotny problem w badaniach statystycznych.

Literatura

Härdle W.: *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Hastie T., Tibshirani R., Friedman J.H.: *The Elements of Statistical Learning*. Springer-Verlag, New York 2001.

Venables W.N.: *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York 1997.

Ranking wydawców za 2002 rok. „Rzeczpospolita" 2003, nr 12, maj.

METHOD OF NONLINEAR SPLINES USED IN BUILDING A REGRESSIVE MODEL

Summary

In this article we describe a nonlinear method of regression which is based on the splines. In this method the regression function is obtained by dividing the domain of X into continuous intervals and representing f by a polinomial in each of them. But we require this function to be smooth. We present two special cases: cubic natural splines and smooth splines. We illustrate this method on the example of market of the publishing houses.