Jacek Stelmach
Uniwersytet Ekonomiczny w Katowicach

O INTERPRETACJI NIEPARAMETRYCZNYCH MODELI REGRESYJNYCH*

Wprowadzenie

Jednym ze sposobów lepszego zrozumienia otaczających nas zjawisk jest budowa ich modelu pozwalająca na uproszczony, ale rzetelny ilościowy lub jakościowy opis reguł charakteryzujących zjawiska. Dobrze skonstruowany model w adekwatny sposób odtwarza badane zjawiska, stanowiąc kompromis między nadmiernym uproszczeniem rzeczywistości a zbytnim nagromadzeniem szczegółów (*Statystyczne metody*…, 1998). Najbardziej znaną metodą budowy wielowymiarowych modeli jest metoda regresji wielorakiej. Jej zaletą jest możliwość interpretacji parametrów modelu regresyjnego, tj. określenie kierunku i siły wpływu zmiennych objaśniających na zmienną objaśnianą. Wadą jest natomiast konieczność spełnienia wymagań:

– homoskedastyczności składnika resztowego,
– normalności rozkładu składnika resztowego,
– braku autokorelacji składnika resztowego,
– niezależności zmiennych objaśniających (Maddala, 2008, s. 165).

W rzeczywistości bardzo często badane zjawiska charakteryzują się:

– nieliniowościami w rzeczywistych procesach,
– zakłóceniami i błędami pomiarowymi,
– korelacjami pomiędzy zmiennymi i ich rozkładami odmiennymi od rozkładu normalnego,
– niestacjonarnością modelowanych procesów,

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

– niewielką liczbą obserwacji oraz występowaniem obserwacji wpływowych i odstających,

co ogranicza możliwości budowy parametrycznych modeli regresyjnych na rzecz modeli nieparametrycznych, których niekwestionowaną zaletą jest brak wymagania znajomości rozkładów cech i postaci analitycznej związku między nimi, a także możliwość tworzenia dokładniejszych prognoz (Gatnar, 2001, s. 16-17). Jednak interpretacja parametrów takich modeli jest niemożliwa bądź bardzo ograniczona. W praktyce interpretacja taka jest najczęściej przydatna w wyspecyfikowanym zakresie zmienności, np. pomiędzy drugim i trzecim kwartylem, albo w zakresie szczególnie interesującym ze względu na specyfikę modelowanego zjawiska lub procesu.

Celem eksperymentu była weryfikacja możliwości interpretacji modeli nieparametrycznych. Istotą proponowanej metody jest utworzenie dodatkowych obserwacji za pomocą zaakceptowanych modeli nieparametrycznych w takim zakresie zmienności, w którym interpretacja parametrów modelu byłaby pożądana. Obserwacje te stanowią próbę wykorzystaną do budowy wtórnego modelu parametrycznego, który można już interpretować. W badaniach porównano właściwości opisanych wyżej wtórnych modeli parametrycznych z modelami parametrycznymi obliczonymi dla próby pierwotnej.

1. Prezentacja hipotezy badawczej

1.1. Prezentacja problemu

Powodem przeprowadzenia eksperymentu była praktyczna potrzeba – konieczność określenia wpływu zmiennych objaśniających na zmienną objaśnianą w pewnym procesie petrochemicznym. Niestety dokładność prognoz modelu parametrycznego nie mieściła się w specyfikacji wymagań (błąd względny MAPE wyniósł 0.14, przy wymaganym poziomie nie większym niż 0.10). Znacznie dokładniejsze były modele nieparametryczne (sieć neuronowa MLP 8-12-1 pozwoliła na prognozy z błędem MAPE równym 0.06, a MAPE prognozy ważonej czterech najlepszych sieci wyniósł 0.05). Nieco gorsze były prognozy modeli obliczonych metodą wektorów nośnych (SVM) – MAPE na poziomie 0.08-0.09. Modele takie pozwalały na podjęcie decyzji zarządczych, ale już nie była możliwa interpretacja wpływu predyktorów na zmienną objaśnianą. Rozpoczęto więc poszukiwanie metody, która pozwoliłaby przynajmniej w przybliżony sposób określić ten wpływ dla regresyjnych modeli nieparametrycznych.

1.2. Postawienie hipotezy

Przyjmując wielowymiarowy model nieparametryczny w postaci $\boldsymbol{y} = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$, gdzie p – liczba wymiarów, możliwe jest ilościowe określenie wpływu zmiennych x_i na zmienną objaśnianą y za pomocą wtórnych modeli parametrycznych w wybranym przedziale zmienności zmiennych objaśniających.

2. Przedstawienie metody

Proponowaną metodę można stosować przy założeniach:

– dopasowanie i prognozy wybranego modelu nieparametrycznego $\boldsymbol{y} = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ są lepsze niż modelu parametrycznego,
– interpretacja modelu nieparametrycznego będzie możliwa w wyspecyfikowanym zakresie zmienności zmiennych objaśniających $(x_{1d}, x_{1g}), \ldots, (x_{pd}, x_{pg})$.

Prezentowana metoda obejmuje poniższą sekwencję:

1. W wybranych zakresach zmienności predyktorów tworzy się próbę wtórną, w której zmienne objaśniające stanowią „kratę”:

$$x_{ij} = x_{id} + j\frac{x_{ig} - x_{id}}{N}, i = 1 \ldots p, j = 1 \ldots N \tag{1}$$

a zmienną objaśnianą oblicza się za pomocą wybranego modelu nieparametrycznego:

$$y_j = f(x_{1j}, . x_{2j}, \ldots, x_{pj}) \tag{2}$$

Łączna liczba obserwacji wtórnej próby wynosi N^p.

2. Dla otrzymanej próby wtórnej tworzy się model parametryczny:

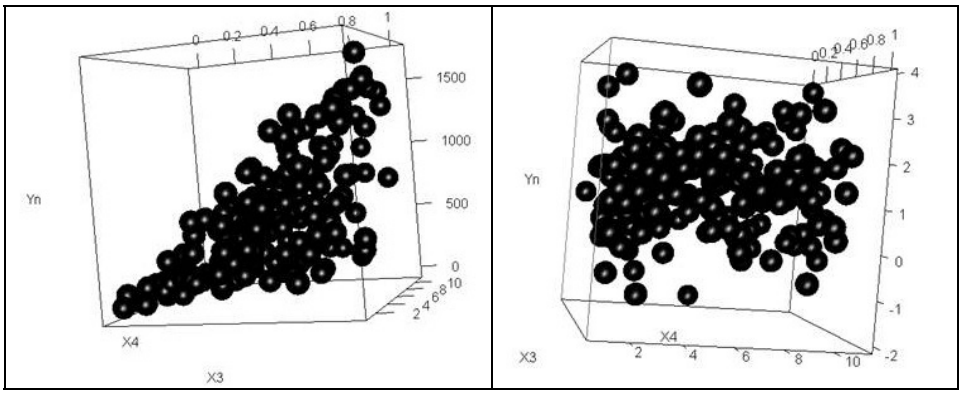
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p \tag{3}$$

3. Parametry β_i pozwalają na określenie wpływu oddziaływania zmiennej objaśniającej x_i na zmienną objaśnianą y.

3. Przeprowadzony eksperyment

Ze względu na poufność danych rzeczywistych przedstawiono wyniki badań przeprowadzonych dla danych empirycznych reprezentowanych przez popularny zestaw *Boston* (506 obserwacji, 13 zmiennych objaśniających), zebrany i opublikowany w 1978 r. przez badaczy zajmujących się zależnością pomiędzy cenami nieruchomości w Bostonie a jakością życia (*Statystyczna analiza danych*…, 2009,

s. 177) oraz dla trzech zestawów danych symulowanych proponowanych przez Friedmana (200 obserwacji) (J. Friedman, 1991, s. 37-44.), symulujących szumy elektroniczne rekomendowane jako zestawy nieliniowe i trudne do wyznaczania modeli regresyjnych. Pseudotrójwymiarowe rzuty pierwszych trzech zmiennych zestawów *Friedman2* oraz *Friedman3* przedstawiono na rysunku 1.



Rys. 1. Rzuty pierwszych trzech zmiennych zestawów: *Friedman2* oraz *Friedman3*

Zestawy te wykorzystywali np. Drucker i in. (1997), badając właściwości modeli obliczonych za pomoca SVM:

– *Friedman1*: $y = 10\sin(\Pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + e_1$

– *Friedman2*: $y = \sqrt{(x_1^2 + (x_2 x_3 - \frac{1}{x_2 x_4})^2)} + e_2$)

– *Friedman3*: $y = \tan^{-1}(\frac{x_2 x_3 - (x_2 x_4)^{-1}}{x_1}) + e_3$

gdzie: x_1, x_2, x_3, x_4 – zmienne o rozkładzie jednostajnym z przedziałów:

$0 < x_1 < 100$; $40\Pi < x_2 < 560\Pi$; $0 < x_3 < 1$; $1 < x_4 < 11$;

e_1, $e_3 \sim N(0,1)$;

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

3.1. Badane metody regresyjne

Badawczy charakter eksperymentu spowodował, że wybrano najbardziej popularne metody nieparametryczne, reprezentujące odmienne podejścia do analizy regresji (szersze omówienie tych metod zob. w *Statystyczna analiza danych*…, 2009, s. 128-259):

– metody oparte na transformacji zmiennych: **metoda rzutowania PPR** (transformacja zmiennych do przestrzeni o mniejszej ilości wymiarów) oraz **metoda addytywna ACE/AREG,**
– **metoda wektorów nośnych SVM,** z automatycznym doborem kluczowych parametrów (typu funkcji jądrowej i parametrów funkcji celu),
– **drzewa regresyjne:** optymalizowane przez przycinanie krawędzi oraz **stochastyczna addytywna metoda drzew regresyjnych MART** z optymalizacją liczby drzew,
– **sieci neuronowe perceptronowe**, jedna warstwa ukryta, dobierane funkcje warstwy ukrytej i wyjściowej, wielkość sieci dobierana automatycznie.

Ocena dokładności dopasowania wyznaczonych modeli regresyjnych została przeprowadzona na podstawie wskaźników:
– ex ante: współczynnik dopasowania $\boldsymbol{R^2}$, błąd średniokwadratowy ***SE***,
– ex post: średni bezwzględny błąd procentowy ***MAPE***, średni bezwzględny błąd ***MAE*** (dla wylosowanych sześciu obserwacji).

3.2. Opis eksperymentu

W eksperymencie utworzono dla każdego z czterech zestawów danych modele parametryczne, wykorzystując metodę Monte Carlo – losowanie przeprowadzono 200 razy, losując z zestawów danych próbę walidacyjną (6 obserwacji) oraz próbę uczącą (pozostałe obserwacje). Wyniki wskaźników ex ante i ex post poddano rangowaniu (rangowanie dla współczynnika dopasowania od wartości największej, a dla błędów – od wartości najmniejszej), a otrzymane rangi uśredniono dla każdej metody, wybierając trzy metody o najmniejszej sumie rang wskaźników: ex ante oraz ex post. Dla każdej wyróżnionej w ten sposób metody regresji wybrano model o najmniejszej sumie rang. Wybrane modele posłużyły do symulacji prób wtórnych o liczebności 10^6 obserwacji w czterech zakresach zmienności (dla każdej zmiennej):
1) minimum – pierwszy kwartyl,
2) pierwszy kwartyl – drugi kwartyl,
3) drugi kwartyl – trzeci kwartyl,
4) trzeci kwartyl – maksimum.

Na podstawie otrzymanych prób utworzono wtórne modele parametryczne (liniowa regresja wieloraka), które porównano z odpowiadającymi im modelami nieparametrycznymi oraz modelami parametrycznymi utworzonymi na podstawie oryginalnych prób.

4. Wyniki eksperymentu

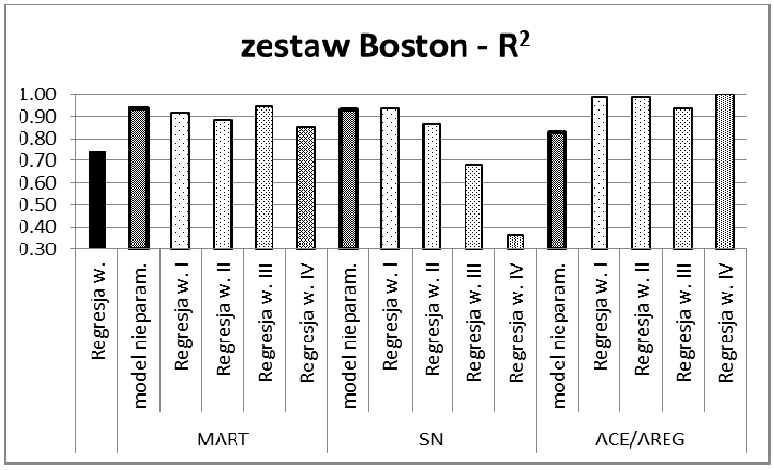
Analiza wskaźników ex ante i ex post modeli nieparametrycznych pozwoliła na wybór tych metod, dla których suma rang tych wskaźników osiągnęła minimum, a następnie wybór już konkretnych modeli nieparametrycznych, gdzie kryterium wyboru była także minimalizacja sumy rang. Wybrane dla każdego zestawu danych modele nieparametryczne uszeregowane zgodnie ze wzrostem sumy rang przedstawiono w tabeli 1.

Tabela 1

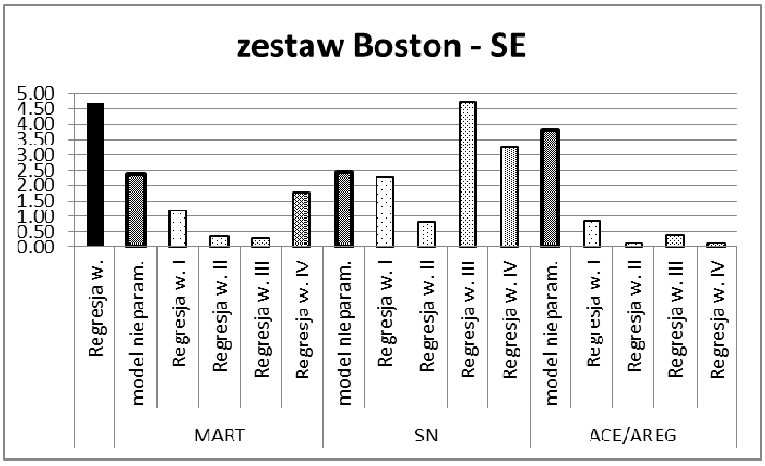
Wybrane modele nieparametryczne

Zestaw danych	Model 1	Model 2	Model 3
Boston	MART	s. neuronowe	ACE/AREG
Friedman1	MART	s. neuronowe	ACE/AREG
Friedman2	s. neuronowe	ACE/AREG	drzewa regresyjne
Friedman3	s. neuronowe	PPR	drzewa regresyjne

Porównanie pierwotnego modelu parametrycznego, wybranych modeli nieparametrycznych oraz odpowiadających im wtórnych modeli parametrycznych obliczonych dla każdego z zakresów zmienności (I, II, III, IV) dla zestawu danych *Boston* przedstawiono na rysunkach 2 i 3.



Rys. 2. Porównanie dopasowania modeli dla zestawu *Boston*



Rys. 3. Porównanie błędu średniokwadratowego modeli dla zestawu *Boston*

Wyniki porównawcze dla wszystkich zestawów danych prezentuje natomiast tabela 2 w porządku: model parametryczny dla oryginalnej próby, wybrany model nieparametryczny i obliczone dla próby otrzymanej z modelu nieparametrycznego – wtórne modele parametryczne.

Tabela 2

Wskaźniki ex ante pierwotnego modelu parametrycznego, wybranych modeli nieparametrycznych oraz wtórnych modeli parametrycznych

Zestaw danych	***Boston***		**Friedman1**		**Friedman2**		**Friedman3**	
Typ modelu	**R^2**	**SE**	**R^2**	**SE**	**R^2**	**SE**	**R^2**	**SE**
Regresja wieloraka	0.74	4.69	0.66	2.65	0.88	129	0.04	1.00
Model 1 nieparametryczny	0.93	2.39	0.95	1.04	0.99	24.4	0.28	0.87
Regresja I	0.91	1.18	0.95	0.28	0.98	4.33	0.84	0.18
Regresja II	0.88	0.37	0.96	0.39	0.99	4.02	0.67	0.09
Regresja III	0.94	0.30	0.85	0.28	0.99	5.98	0.73	0.01
Regresja IV	0.85	1.77	0.92	0.22	0.99	8.19	0.77	0.01
Model 2 nieparametryczny	0.93	2.46	0.95	0.96	0.99	0.19	0.50	0.72
Regresja I	0.94	2.28	0.94	0.23	0.99	0.02	0.25	0.50
Regresja II	0.87	0.82	0.98	0.24	0.99	0.03	0.13	0.41
Regresja III	0.68	4.71	0.85	0.28	0.99	0.01	0.41	0.57
Regresja IV	0.36	3.26	0.99	0.09	0.99	0.01	0.41	0.30
Model 3 nieparametryczny	0.83	3.80	0.89	1.47	0.96	80.7	0.40	0.79
Regresja I	0.99	0.86	0.99	0.02	0.50	18.5	X	X
Regresja II	0.98	0.11	0.99	0.11	0.79	40.0	X	X
Regresja III	0.93	0.39	0.97	0.11	0.66	95.2	X	X
Regresja IV	0.99	0.13	0.99	0.06	0.52	71.3	X	X

Należy podkreślić, że dla zestawu danych *Friedman3* trudno było utworzyć zadowalające modele. Analiza modelu parametrycznego wskazała na bardzo niską istotność statystyczną (*p-value F* = 0.06), wtórne modele parametryczne

dla modelu drzew regresyjnych niemożliwe do wyznaczenia ze względu na źle uwarunkowane macierze.

We wszystkich przypadkach wskaźniki ex ante wskazywały na znacznie lepsze dopasowanie zarówno modeli nieparametrycznych, jak i wtórnych modeli parametrycznych – w porównaniu z modelami parametrycznymi otrzymanymi dla oryginalnych prób.

Podsumowanie

Przeprowadzony eksperyment potwierdza możliwość interpretacji modeli nieparametrycznych w wybranym zakresie zmienności zmiennych objaśnianych. Warunkiem jest dobre dopasowanie tych modeli. Interpretacja zgodnie z zaproponowaną metodą nie powinna być stosowana, jeśli modele parametryczne nie są mniej dokładne od modeli nieparametrycznych oraz jeśli nie uda się utworzyć modeli nieparametrycznych o zadowalającej badacza jakości, co ogranicza wiarygodność samej interpretacji.

Literatura

Drucker C.J., Burges C.J.C., Kaufman L., Smola A., Vapnik V. (1997): *Support Vector Regression Machines*. „Advances in Neural Information Processing Systems", Vol. 9.

Friedman J. (1991): *Multivariate Adaptive Regression Splines.* „Annals of Statistics", Vol. 19, Institute of Mathematical Statistics, Stanford University.

Gatnar A. (2001): *Nieparametryczna metoda dyskryminacji i regresji.* Wydawnictwo Naukowe PWN, Warszawa.

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

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

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

Statystyczne metody analizy danych (1998). Red. W. Ostasiewicz. Wydawnictwo AE, Wrocław.

Tadeusiewicz R., Lula P. (2000): *Neuronowe metody analizy szeregów czasowych i możliwości ich zastosowań w zagadnieniach biomedycznych.* W: *Biocybernetyka i inżynieria biomedyczna*. *Tom 6*. *Sieci neuronowe*. Red. M. Nałęcz. Akademicka Oficyna Wydawnicza Exit, Warszawa.

PARAMETRIC INTERPRETATION OF NON-PARAMETRIC REGRESSION MODELS

Summary

The advantage of the parametric regression models is the possibility of interpretation of the parameters of the regression model, i.e. to determine the direction and strength of the influence of predictors on the dependent variable. Unfortunately, in practice – the non-linearity of the real processes, the influence of the phenomena with various probability distributions and a small number of observations limits the building of parametric models while the interpretation of non-parametric models is either impossible or very limited.

Frequently such interpretation is useful in the specified range of variation. This may be a typical range of variation – for example, between the second and third quartiles, or a specific range due to the nature of the modeled phenomenon or process. It is difficult however, to build parametric models based only on the range of explanatory variables, because in this way we exclude observations giving additional knowledge into the model.

The essence of this study is to enable the interpretation of non-parametric models through the creation of additional observations with these models in an interesting range of explanatory variables. These observations create secondary dataset used for the construction of a parametric model, which can now be interpreted. Presented investigations compare – using simulation – parametric models created for secondary sample with parametric models calculated for the original data.