Joanna Trzęsiok

Uniwersytet Ekonomiczny w Katowicach

WYKORZYSTANIE REGRESJI NIEPARAMETRYCZNEJ DO MODELOWANIA WIELKOŚCI OSZCZĘDNOŚCI GOSPODARSTW DOMOWYCH

Wprowadzenie

Nieparametryczne metody regresji można zdefiniować jako takie, w których postać modelu nie jest jednoznacznie określona, w tym sensie, że występuje przynajmniej jeden z poniższych przypadków:

- nie jest ściśle zadana postać analityczna funkcji składowych modelu,
- liczba funkcji składowych modelu nie jest z góry ustalona,
- na etapie budowy modelu nie jest jednoznacznie określony zestaw zmiennych, który zostanie uwzględniony w modelu końcowym.

Ponadto, w modelach nieparametrycznych nie zachodzi konieczność testowania normalności rozkładu składnika losowego, czy sprawdzania współliniowości zmiennych objaśniających.

Metody nieparametryczne stanowią podejście alternatywne w stosunku do klasycznej metody regresji wielorakiej, ponieważ zostały skonstruowane tak, by rozwiązywały zadania regresji, kiedy choć część restrykcyjnych założeń klasycznego modelu liniowego nie jest spełniona. W związku z tym modele nieparametryczne charakteryzują się dużo większą elastycznością, a dodatkowo zakres ich potencjalnych zastosowań jest znacznie szerszy.

Nieparametryczne metody regresji są zróżnicowaną i dynamicznie rozwijającą się grupą metod statystycznej analizy danych. Decydującym czynnikiem, który wpłynął na ich rozwój był postęp technologii informatycznych, który umożliwił budowę modeli z wykorzystaniem złożonych algorytmów numerycznych.

W tym artykule przedstawiono wykorzystanie regresji nieparametrycznej do modelowania wielkości oszczędności gospodarstw domowych. Analizę przeprowadzono na danych rzeczywistych, pochodzących z badania przeprowadzonego w 2000 r. przez Główny Urząd Statystyczny [GUS, 2001].

Przeprowadzone badania porównawcze pokazują, że niemożliwe jest wskazanie najlepszej metody regresji, która w każdej sytuacji, niezależnie od rozważanego zbioru danych, dawałaby najniższe błędy predykcji [Meyer et al., 2003]. W przypadku szacowania wielkości oszczędności gospodarstw domowych nie było też merytorycznych argumentów wspomagających wybór metody regresji. Zastosowano więc procedurę polegającą na zbudowaniu kilku modeli regresji nieparametrycznych i wybrano model o najlepszej zdolności predykcji, który następnie poprawiono poprzez wyeliminowanie z niego zmiennych nieistotnych. Wszystkie obliczenia i analizy wykonano z wykorzystaniem programu statystycznego **R**.

1. Analizowany zbiór danych

W analizie, mającej na celu modelowanie wielkości oszczędności gospodarstw domowych, wykorzystano dane uzyskane z badania przeprowadzonego metodą reprezentacyjną przez GUS w roku 2000. Było to badanie budżetów gospodarstw domowych, które spełnia ważną rolę w analizach poziomu życia Polaków, ponieważ jest źródłem informacji m.in. o przychodach, rozchodach czy spożyciu ilościowym żywności dla różnych grup ludności. Otrzymywane wyniki są wykorzystywane do opracowań prognostycznych oraz analiz ekonomicznych.

W niniejszym artykule modelowano wielkość oszczędności na podstawie danych dotyczących gospodarstw domowych pracowników, tj. gospodarstw, w których wyłącznym lub głównym źródłem utrzymania był dochód z pracy najemnej w sektorze publicznym lub prywatnym.

Krótką charakterystykę analizowanego zbioru danych, który na potrzeby tej pracy nazwano *Budżety*, przedstawiono w tabeli 1.

Tabela 1

Charakterystyki zbioru danych *Budżety*

Liczebność zbioru	Liczba zmiennych objaśniających		
	ilorazowych	porządkowych	nominalnych
14 423	3	3	1

Zmiennymi objaśniającymi w zbiorze *Budżety* są:

doch – miesięczny dochód gospodarstwa domowego,
wydg – miesięczne wydatki gospodarstwa domowego,
klm – klasa miejscowości (wyróżniono miasta o liczbie mieszkańców: powyżej 500 tys., 200-500 tys., 100-200 tys., 20- 100 tys., oraz wieś),
trb – typ rodziny biologicznej (wyróżniono małżeństwa: bez dzieci, z jednym dzieckiem na utrzymaniu, z 2 dzieci na utrzymaniu, z 3 dzieci na utrzymaniu, z 4 i więcej dzieci na utrzymaniu, samotnych rodziców z dziećmi na utrzymaniu oraz pozostałe),
ocdoch – subiektywna ocena dochodów gospodarstwa, która jest związana z odpowiedzią na pytanie ankietowe „czy gospodarstwo wiąże koniec z końcem?”,
przynpie – wartość przychodów niepieniężnych gospodarstwa, pochodzących np. z darowizn lub niepieniężnej pomocy społecznej,
wggd – wykształcenie głowy gospodarstwa domowego (wyróżniono osoby: bez wykształcenia, z wykształceniem podstawowym, zasadniczym zawodowym, średnim oraz wyższym),

zaś zmienną zależną jest:

oszcz – wielkość oszczędności gospodarstwa domowego [w zł].

W trakcie przygotowywania zbioru do analizy wykryto dwie obserwacje odstające, które usunięto ze zbioru *Budżety*. Jedna z tych obserwacji zawierała ujemną wielkość oszczędności, a druga – oszczędności równe 210 tys. zł, różniła się od mediany o ponad 500 odchyleń ćwiartkowych. Ostatecznie w badaniu wykorzystano zbiór złożony z 14 423 obserwacji.

W tabeli 2 przedstawiono wybrane statystyki opisowe dla zmiennej zależnej *oszcz*. Zaobserwowano bardzo silne zróżnicowanie zmiennej zależnej, jak i silną asymetrię prawostronną (zestandaryzowany moment centralny trzeciego rzędu równy 9,2). Wyklucza to modelowanie wielkości oszczędności z wykorzystaniem klasycznej, liniowej metody regresji wielorakiej.

Tabela 2

Charakterystyki opisowe zmiennej zależnej *oszcz* w zbiorze danych *Budżety*

Średnia	Odchylenie standardowe	Współczynnik asymetrii
735,5 zł	1 054,2 zł	9,2
Minimum	Mediana	Maksimum
0 zł	455 zł	38 480 zł

2. Metody regresji wykorzystane w analizie

Wybór najlepszej metody regresji do rozwiązania zadanego problemu jest dylematem, z którym spotkało się wielu badaczy. Tak jak już wspomniano we wprowadzeniu, niemożliwe jest wskazanie najlepszej metody, która niezależnie od rozważanego zbioru danych, generuje modele o najmniejszych błędach średniokwadratowych. W tej analizie także charakter badanego zbioru danych nie determinuje wyboru odpowiedniej metody. Zaproponowano więc zastosowanie procedury, która polegała na zbudowaniu wielu modeli regresji (dla różnych wartości parametrów) i wyborze najlepszego z nich (pod względem dokładności predykcji), który to model został następnie poprawiony poprzez wyeliminowanie z zestawu zmiennych objaśniających, tych cech, które nie wpływały istotnie na wielkość oszczędności (*oszcz*).

Jak wspomniano, w pierwszym etapie badania zbudowano wiele modeli, za pomocą następujących nieparametrycznych metod regresji*:

- metody rzutowania PPR [Friedman, Stuetzle, 1981],
- metody ACE polegającej na jednoczesnej transformacji wszystkich zmiennych [Breiman, Friedman, 1985],
- metody rekurencyjnego podziału – RPART [Breiman et al., 1984],
- metody polegającej na równoległym łączeniu drzew regresyjnych [Breiman, 1996] (oznaczonej jako BAGGING),
- stochastycznej, addytywnej metody drzew regresyjnych MART [Friedman, 1999a, 1999b],
- metody zagregowanych drzew regresyjnych Breimana – RANDOM FORESTS [Breiman, 2001],
- wielowymiarowej metody krzywych sklejanych POLYMARS [Kooperberg et al., 1997],
- metody wektorów nośnych SVM [Vapnik, 1998],
- metody wykorzystującej sieci neuronowe (oznaczonej jako NNET) [Bishop, 1995].

Dla każdej z wymienionych metod zbudowano (wykorzystując odpowiednie funkcje programu statystycznego **R**) modele dla różnych zestawów parametrów. Jednak w ostatecznym zestawieniu daną metodę reprezentuje tylko jeden model – ten w którym wykorzystano optymalną konfigurację wartości parametrów (dającą najmniejsze wartości błędu średniokwadratowego). Zwieńczeniem

* Wymienione nieparametryczne metody regresji były przedmiotem badań autorki w poprzednich latach, których wyniki zostały opisane w innych publikacjach. W tym miejscu ograniczono się jedynie do podania artykułów źródłowych, w których można znaleźć szerszą charakterystykę tych metod.

tego etapu procedury badawczej jest stworzenie rankingu modeli (tabela 3), pod względem dokładności predykcji, ocenianej za pomocą estymatora punktowego, jakim jest błąd średniokwadratowy, który został obliczony metodą sprawdzania krzyżowego (MSE_{CV}).

Tabela 3

Błędy średniokwadratowe MSE_{CV} obliczone dla modeli otrzymanych różnymi metodami regresji, dla zbioru *Budżety*

Metoda	Błąd MSE_{CV}
MART	732 410
PPR	738 286
R.FOREST	742 864
ACE	745 925
NNET	759 690
POLYMARS	765 690
SVM	777 784
BAGGING	803 413
RPART	821 577
Metoda	Błąd MSE_{CV}

W tym przypadku, najlepszym pod względem zdolności predykcyjnych jest model zbudowany addytywną metodą drzew regresyjnych MART. Charakteryzuje się on najniższym błędem średniokwadratowym obliczonym metodą sprawdzania krzyżowego (MSE_{CV}). Model ten został wykorzystany w dalszej analizie.

3. Identyfikacja zmiennych istotnych i nieistotnych

Jedną z największych wad nieparametrycznych metod regresji jest to, że większość z nich działa na zasadzie „czarnej skrzynki". Wyjątek stanowi metoda rekurencyjnego podziału, dla której Breiman zaproponował miernik oceny siły wpływu każdej ze zmiennych objaśniających na zmienną zależną [Breiman et al., 1984]. Dla zmiennej X_j miarę tę można przedstawić w postaci:

$$W_T(X_j) = \sqrt{\sum_{p=1}^{P-1} \varphi_p^2 \cdot I\left(\upsilon(p) = j\right)}, \tag{1}$$

gdzie: T to model drzewa regresyjnego, P – liczba węzłów* tego drzewa, $\upsilon(p)$ to numer zmiennej objaśniającej występującej w węźle p, zaś współczynnik φ_p^2 dany jest wzorem:

* Węzeł reprezentuje w graficznej postaci drzewa (grafie) podział segmentu na dwa podzbiory [Gatnar, 2001].

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

Wartości współczynnika φ_p^2, zsumowane po wszystkich węzłach p, w których występuje zmienna objaśniająca X_j, reprezentują wpływ tej zmiennej na Y.

Wzór (1) może zostać w prosty sposób uogólniony i zastosowany dla modeli zagregowanych drzew regresyjnych, w tym również dla modeli uzyskanych metodą MART [Hastie et al., 2001, s. 332]. W modelach tych siłę wpływu każdej ze zmiennych X_j na zmienną zależną Y określa miara

$$W(X_j) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(W_{T_k}(X_j)\right)^2}, \qquad (3)$$

gdzie T_k jest modelem składowym – pojedynczym drzewem regresyjnym (dla $k = 1, \ldots, K$).

Funkcja gbm, z biblioteki gbm programu statystycznego **R**, pozwala na obliczenie miary przedstawionej wzorem (3) i zbudowanie rankingu zmiennych objaśniających pod względem ich wpływu na zmienną zależną. Ranking ten, wraz z procentowym oszacowaniem wpływu zmiennych objaśniających na zmienną *oszcz*, przedstawiono w tabeli 4.

Tabela 4

Ranking zmiennych objaśniających pod względem ich siły wpływu na zmienną *oszcz*, dla metody MART

Nr	Zmienna objaśniająca	Siła wpływu na zmienną *oszcz*
1	*Doch*	77,7 %
2	*Wydg*	10,4 %
3	*Przynpie*	4,5 %
4	*Ocdoch*	3,6 %
5	*Klm*	2,3 %
6	*Wggd*	1,0 %
7	*Trb*	0,4 %

Przedstawiony ranking zmiennych pokazuje, iż największy wpływ na wielkość oszczędności ma dochód gospodarstwa domowego. Relatywnie silną zależność obserwuje się również pomiędzy zmienną *oszcz* a wydatkami badanych gospodarstw. Można powiedzieć, że jest to wynik zgodny z oczekiwaniami badacza i teorią ekonomii. Najmniejszy wpływ na wielkość oszczędności ma zmienna przedstawiająca typ rodziny biologicznej.

W dalszej części analizy starano się wyodrębnić zmienne istotnie wpływające na zmienną *oszcz* i tylko te wprowadzić do modelu, zbudowanego metodą MART. Pozwoliło to na uzyskanie dodatkowych informacji na temat zależności między badanymi cechami. Ponadto, zredukowanie liczby zmiennych prowadzi najczęściej do zmniejszenia złożoności modelu.

Jak wspomniano, funkcja gbm pozwala na stworzenie rankingu zmiennych objaśniających, ze względu na siłę ich wpływu na zmienną zależną, jednak nie oddziela ona zmiennych istotnie od nieistotnie wpływających na zmienną *oszcz*. Rozdzielenie zmiennych objaśniających (na istotne i nieistotne) poprzez ustalenie odpowiedniego poziomu siły wpływu zmiennych w rankingu (w tabeli 4) miałoby charakter subiektywny. Ponadto zidentyfikowanie zmiennych istotnych powinno uwzględniać interakcje między cechami, a nie tylko wpływ na zmienną zależną każdej zmiennej objaśniającej z osobna.

W celu wyodrębnienia zestawu zmiennych istotnie wpływających na wielkość oszczędności, przeprowadzono procedurę eliminacji zmiennych blokiem [Nagatani, Abe, 2007; Trzęsiok, 2010]. W tablicy 5 przedstawiono szczegółowo kroki algorytmu zastosowanej metody eliminacji cech.

Tabela 5

Algorytm metody eliminacji zmiennych blokiem

Krok 1	Zbuduj model regresyjny na zbiorze uczącym D wykorzystując pełen zestaw zmiennych. Oblicz błąd średniokwadratowy tego modelu $MSE_{CV}(D)$ metodą sprawdzania krzyżowego. Utwórz pomocniczy zbiór S będący kopią zbioru D
Krok 2	Poprzez wyłączenie tymczasowo ze zbioru S kolejno każdej ze zmiennych wygeneruj wiele zmodyfikowanych zbiorów uczących na bazie S. Zbuduj na tak zmodyfikowanych zbiorach modele regresyjne
Krok 3	Dla każdego modelu z kroku 2 oblicz metodą sprawdzania krzyżowego błąd średniokwadratowy MSE_{CV}
Krok 4	Zidentyfikuj wszystkie modele z wyłączoną zmienną, dla których wartość błędu MSE_{CV} jest mniejsza niż wartość $MSE_{CV}(D)$. Jeśli warunek nie jest spełniony dla żadnego modelu, to uznaj wszystkie zmienne za istotne i zakończ procedurę
Krok 5	Usuń tymczasowo ze zbioru S wszystkie zmienne zidentyfikowane w kroku 4. Zbuduj na tym zbiorze nowy model i oblicz jego błąd średniokwadratowy. Jeśli obliczony błąd jest mniejszy od wartości $MSE_{CV}(D)$, to zapamiętaj tak zredukowany zbiór S i powróć do kroku 2
Krok 6	W przeciwnym przypadku przywróć zbiór S z kroku 2 i zastosuj algorytm połowienia do zidentyfikowania mniej licznego bloku zmiennych do usunięcia: a) uporządkuj modele z kroku 4 według rosnących wartości MSE_{CV}, b) tymczasowo usuń ze zbioru S pierwszą połowę zmiennych, odpowiadającą uporządkowanym w kroku 6a) modelom. Zbuduj model regresyjny na zbiorze S, z którego tymczasowo usunięto blok zmiennych o połowę mniej liczny niż uprzednio, i oblicz błąd MSE_{CV}. Jeśli obliczony błąd jest mniejszy od $MSE_{CV}(D)$, to pozostaw tak zredukowany zbiór S i powróć do kroku 2. W przeciwnym przypadku przywróć zbiór S z kroku 2 i rekurencyjnie zastosuj metodę połowienia dla mniej licznego bloku zmiennych (przejdź do kroku 6b).

Wyniki tej procedury przeprowadzonej na zbiorze *Budżety* dla metody MART przedstawiono w tabeli 6.

Tabela 6

Wynik działania procedury eliminacji zmiennych blokiem dla zbioru *Budżety*, dla metody MART

Etap	Usunięte zmienne	Nazwy usuniętych zmiennych	Błąd MSE_{CV} modelu
0	Ø		732 410
1	4, 5	*trb*, *ocdoch*	730 262
2	3, 6	*klm*, *przynpie*	731 077
3	1, 2, 7	*Doch*, *wydg*, *wggd*	

Wyniki przedstawione w tabeli 6 pokazują, iż eliminacja zmiennych blokiem wskazuje na trzy zmienne objaśniające istotnie wpływające na wielkość oszczędności. Są to: dochody i wydatki gospodarstwa domowego oraz wykształcenie głowy gospodarstwa domowego. Zauważyć można, że zmienna *wggd* w przedstawionym rankingu (tabela 4) nie charakteryzowała się zbyt silnym wpływem na zmienną *oszcz*, jednak w eliminacji zmiennych blokiem oceniany jest wpływ grup zmiennych, a nie tylko pojedynczych cech. Uwzględniane są więc również interakcje pomiędzy zmiennymi, zatem zmienna określająca wykształcenie głowy gospodarstwa *wggd*, pomimo słabego (indywidualnie) wpływu na zmienną zależną, tworzy w interakcji ze zmiennymi przedstawiającymi dochody (*doch*) oraz wydatki (*wydg*) grupę cech istotnie oddziaływujących na wielkość modelowanych oszczędności w badanych gospodarstwach domowych.

Model regresji otrzymany po zastosowaniu metody eliminacji zmiennych blokiem ma znacznie mniejszą złożoność, a jednocześnie charakteryzuje się niższym błędem średniokwadratowym MSE_{CV} = 731 077 niż model zbudowany na całym zestawie cech $MSE_{CV}(D)$ = 732 410.

Końcowy model MART zbudowano z 9 672 modeli składowych. Metoda MART jako metoda agregacji pojedynczych funkcji składowych, nie pozwala niestety na wyznaczenie i interpretowanie parametrów modelu.

Podsumowanie

W artykule przedstawiono procedurę badawczą, której zastosowanie prowadzi do wyboru optymalnego nieparametrycznego modelu regresji, wykorzystanego do analizy zbioru danych *Budżety*.

W pierwszym kroku tej procedury zbudowanych zostało wiele modeli regresji, dla różnych zestawów parametrów. Spośród nich wybrano model uzyska-

ny za pomocą metody MART, ponieważ charakteryzował się on najlepszymi zdolnościami predykcyjnymi, mierzonymi wielkością błędu średniokwadratowego obliczonego metodą sprawdzania krzyżowego (MSE_{CV}).

W drugim etapie przeprowadzono dodatkowo procedurę eliminacji zmiennych blokiem i do modelu MART wprowadzono tylko te zmienne, które istotnie wpływały na zmienną zależną. Ostatecznie, uzyskany model optymalny opisywał zależność wielkości oszczędności tylko od trzech zmiennych objaśniających: dochodów i wydatków gospodarstwa domowego oraz wykształcenia głowy gospodarstwa domowego. Ponadto wartość błędu MSE_{CV} tego modelu była niższa od wartości błędu średniokwadratowego modelu zbudowanego na całym zestawie zmiennych objaśniających. Model optymalny charakteryzował się również mniejszą złożonością.

Wykorzystanie zaproponowanej metody pozyskiwania modelu optymalnego, do analizy postawionego zadania regresji, jest rekomendowane szczególnie wtedy, gdy badacz nie ma dodatkowych argumentów przemawiających za wyborem określonej metody.

Literatura

Bishop C. (1995): Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Breiman L. (1996): Bagging Predictors. „Machine Learning", 24, s. 123-140.

Breiman L. (2001): Random Forests. „Machine Learning", 45, s. 5-32.

Breiman L., Friedman J.H. (1985): Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion). „Journal of the American Statistical Association", 80, s. 580-619.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984): Classification and Regression Trees. Chapman & Hall, New York.

Friedman J. (1999a): Greedy Function Approximation: A Gradient Boosting Machine. Technical Report. Department of Statistics, Stanford University, Redwood City, CA.

Friedman J. (1999b): Stochastic Gradient Boosting. Technical Report. Stanford University, Dept. of Statistics.

Friedman J., Stuetzle W. (1981): Projection Pursuit Regression. „Journal of the American Statistical Association", 76, s. 817-823.

Gatnar E. (2001): Nieparametryczna metoda dyskryminacji i regresji. „Biblioteka ekonometryczna", Wydawnictwo Naukowe PWN, Warszawa.

GUS (2001): Warunki życia ludności w 2000 r. Warszawa.

Hastie T., Tibshirani R., Friedman J. (2001): The Elements of Statistical Learning: Data Mining, Inference and Prediction. „Springer Series in Statistics", Springer Verlag, New York.

Kooperberg C., Bose S., Stone C. (1997): Polychotomous Regression. „Journal of the American Statistical Association", 92, s. 117-127.

Meyer D., Leisch F., Hornik K. (2003): The Support Vector Machine under Test. „Neurocomputing", 55(1-2), s. 169-186.

Nagatani T., Abe S. (2007): Backward Variable Selection of Support Vector Regressors by Block Deletion. Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007, IEEE, s. 2117-2122.

Trzęsiok J. (2010): Dobór zmiennych do modelu regresyjnego zbudowanego za pomocą wybranych nieparametrycznych metod regresji. W: Klasyfikacja i analiza danych – teoria i zastosowania. Red. K. Jajuga, M. Walesiak. Taksonomia, 17, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 172-180.

Vapnik V. (1998): Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, New York.

NONPARAMETRIC REGRESSION APPLIED TO MODELLING HOUSEHOLD SAVINGS

Summary

In the paper the procedure for selecting the best nonparametric model for a given problem of regression is presented. This procedure has two stages. In the first one, many nonparametric models of regression, for different parameters settings, are built. Then the model with the smallest mean squared error is chosen. In the second stage, the method for the reduction of insignificant predictors is used. This procedure is applied to modelling household savings.