Joanna Trzęsiok

Uniwersytet Ekonomiczny w Katowicach
Katedra Matematyki
joanna.trzesiok@ue.katowice.pl

PORÓWNANIE ZDOLNOŚCI PREDYKCYJNYCH MODELU REGRESJI GRZBIETOWEJ Z WYBRANYMI NIEPARAMETRYCZNYMI MODELAMI REGRESJI

Wprowadzenie

Wobec coraz powszechniejszej informatyzacji życia gospodarczego, ilość informacji gromadzonych i przechowywanych w bazach danych wzrosła gwałtownie, prowadząc do jej nadmiarowości. Wpływa to na konieczność stosowania w analizach coraz lepszych metod statystycznych, które muszą być adekwatne do poziomu złożoności badanych zjawisk.

Najbardziej znanym i często stosowanym modelem regresji jest regresja wieloraka (o postaci liniowej), której parametry szacuje się metodą najmniejszych kwadratów. Wśród zalet tej metody należy wymienić prostotę oraz łatwą interpretowalność parametrów otrzymanego modelu. Jej wadą są restrykcyjne założenia nałożone na zmienne charakteryzujące analizowane zjawiska. Ograniczenia te często wykluczają możliwość stosowania w praktyce tej metody w klasycznej postaci.

W związku z tym pojawiła się potrzeba stosowania innych narzędzi analizy regresji, które nakładałyby mniej założeń na badane zjawiska i tym samym były przydatne do rozwiązywania tych problemów, do których nie można zastosować liniowego modelu regresji wielorakiej. Jedną z takich metod jest **regresja grzbietowa** (*ridge regression*), zaproponowana przez Hoerla i Kennarda [6], [7]. Metoda ta poprzez wprowadzenie do modelu pewnej stałej, rozwiązuje problem współliniowości zmiennych objaśniających, jak również redukuje ich liczbę[1].

[1] Metoda ta zostanie szerzej przedstawiona w dalszej części artykułu.

Innymi skutecznymi metodami wielowymiarowej analizy danych są **nieparametryczne metody regresji**, które 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.

Celem artykułu jest porównanie modeli otrzymywanych za pomocą regresji grzbietowej z wybranymi nieparametrycznymi metodami regresji, pod względem zdolności predykcyjnych, które w tej pracy będą rozumiane jako ocena, na ile wartości teoretyczne, oszacowane na podstawie zbudowanego modelu, różnią się od wartości rzeczywistych dla obserwacji spoza zbioru uczącego. W zestawieniach różnych metod regresji, pod względem zdolności predykcji, liniowy model najczęściej zajmuje ostatnie miejsce [12]. Interesujące wydaje się więc przeprowadzenie badań porównawczych dla regresji grzbietowej, jako metody będącej „ulepszoną wersją" liniowej regresji wielorakiej.

Ze względu na charakter nieparametrycznych metod regresji – ich odmienne mechanizmy działania, niemożliwe jest analityczne porównanie otrzymywanych modeli. Z tego względu badania porównawcze przeprowadzono za pomocą procedur symulacyjnych, na zbiorach danych standardowo wykorzystywanych do badania własności różnych metod regresji. Wszystkie analizy i obliczenia przeprowadzono z wykorzystaniem programu statystycznego **R** z dołączonymi bibliotekami tego programu[2].

1. Regresja grzbietowa

W modelu regresji wielorakiej o postaci liniowej zależność zmiennej Y od zmiennych objaśniających $X_1, \ldots, X_m$ można przedstawić jako:

$$Y = a_0 + \sum_{j=1}^{m} a_j X_j + \varepsilon , \tag{1}$$

[2] Program statystyczny **R** jest produktem darmowym, który jest dostępny wraz z dodatkowymi bibliotekami pod adresem http://www.r-project.org.

gdzie a_j (dla $j = 0,1,\ldots,m$) to parametry strukturalne modelu. Równanie (1) można równoważnie zapisać w postaci macierzowej

$$\mathbf{y} = \mathbf{X}'\mathbf{a} + \boldsymbol{\varepsilon}, \tag{2}$$

gdzie

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X}' = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{21} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{mn} \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Do oszacowania wartości parametrów strukturalnych modelu regresji wielorakiej (2) stosuje się metodę najmniejszych kwadratów, w której jest rozwiązywany problem optymalizacyjny:

$$\sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{m} a_j x_j \right)^2 \to \min, \tag{3}$$

który również można zapisać w postaci macierzowej:

$$\left(\mathbf{y} - \mathbf{X}'\mathbf{a}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}'\mathbf{a}\right) \to \min. \tag{4}$$

Rozwiązaniem tak postawionego zadania minimalizacji (4) jest estymator wektora **a**:

$$\hat{\mathbf{a}} = \left(\mathbf{X}'^{\mathrm{T}}\mathbf{X}'\right)^{-1}\mathbf{X}'^{\mathrm{T}}\mathbf{y}. \tag{5}$$

Jednak w przypadku skorelowania zmiennych objaśniających niemożliwe jest odwrócenie macierzy $\mathbf{X}'^{\mathrm{T}} \cdot \mathbf{X}'$ i tym samym wyznaczenie estymatora $\hat{\mathbf{a}}$. Problem ten został rozwiązany w 1977 r. przez Hoerla i Kennarda, którzy jako pierwsi zastosowali **regresję grzbietową**. Ideą tej metody jest przezwyciężenie problemu osobliwości macierzy $\mathbf{X}'^{\mathrm{T}} \cdot \mathbf{X}'$ poprzez dodanie do jej przekątnej, przed odwróceniem, stałej, dodatniej wartości λ.

Analitycznie zagadnienie to zapisuje się w postaci zadania minimalizacji:

$$\sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{m} a_j x_j \right)^2 + \lambda\sum_{j=1}^{m} a_j^2 \to \min, \tag{6}$$

gdzie: $\lambda > 0$ – parametr kary. Jeśli $\lambda = 0$, to estymatory parametrów strukturalnych są po prostu wyznaczone za pomocą metody najmniejszych kwadratów. Jeśli $\lambda \to \infty$, to uzyskany model regresji będzie złożony tylko z wyrazu wolnego.

Hoerl i Kennard w swojej pracy proponowali testowanie różnych wartości λ i wybór tej z nich, dla której „układ się ustabilizuje". Ze względu na różne zakresy zmienności cech do przeprowadzenia poprawnej estymacji parametrów modelu regresji grzbietowej, wszystkie zmienne powinny zostać zestandaryzowane.

Łatwo zauważyć, że w zadaniu (6) nie nakłada się kary na wyraz wolny modelu. Jego estymator wyznacza się ze wzoru:

$$a_0 = \frac{1}{n}\sum_{i=1}^{n} y_i \,. \tag{7}$$

W związku z tym zadanie minimalizacji (6) można przedstawić w postaci macierzowej:

$$\left(\mathbf{y} - \mathbf{X}\mathbf{a}'\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{a}'\right) + \lambda\,\mathbf{a}'^{\mathrm{T}}\mathbf{a}' \rightarrow \min\,, \tag{8}$$

gdzie do modelu wprowadzono zmienne po standaryzacji, natomiast $\mathbf{a}'$ to wektor parametrów strukturalnych $\mathbf{a}' = [a_1, \ldots, a_m]^{\mathrm{T}}$. Rozwiązanie tak postawionego problemu minimalizacji prezentuje wzór [5, s. 60]:

$$\hat{\mathbf{a}}' = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\,, \tag{9}$$

gdzie $\mathbf{X}$ jest macierzą realizacji zmiennych objaśniających, po zastosowaniu formuły standaryzacyjnej (bez jedynek w pierwszej kolumnie).

W praktyce sytuacja dokładnej współliniowości (gdy macierz $\mathbf{X}'^{\mathrm{T}} \cdot \mathbf{X}'$ jest osobliwa) występuje rzadko. Najczęściej występuje zjawisko *przybliżonej* współliniowości, które ma niekorzystne konsekwencje, np.:

- niemożliwy staje się prawidłowy pomiar siły oddziaływania zmiennych objaśniających na zmienną zależną,
- oceny wariancji oszacowanych estymatorów (5), odpowiadające skorelowanym zmiennym, są bardzo wysokie,
- oszacowania parametrów są bardzo wrażliwe nawet na niewielkie zmiany liczby obserwacji.

Skorelowanie zmiennych objaśniających samo w sobie nie zawsze jest powodem problemów przy wnioskowaniu [11]. Współliniowość nie powoduje utraty własności nieobciążoności oraz efektywności estymatorów (5), jednak ze względu na ich dużą wariancję, można się spodziewać niewielkiej precyzji ocen parametrów [13]. W tym kontekście dobór odpowiedniej wartości parametru λ (parametru kary) w regresji grzbietowej jest również próbą znalezienia kompro-

misu pomiędzy obciążeniem a wariancją w przypadku przybliżonej współliniowości zmiennych objaśniających.

Można także zauważyć, że problem minimalizacji (6) jest typowym przykładem zagadnienia regularyzacji [5, s. 34]:

$$PRSS(f,\lambda) = RSS(f) + \lambda J(f) \to \min, \tag{10}$$

gdzie:

$RSS(f)$ – miara jakości dopasowania funkcji regresji (obliczona na zbiorze uczącym),

$J(f)$ – stopień złożoności modelu,

parametr λ – kompromis pomiędzy dopasowaniem modelu a jego złożonością, określając proporcje pomiędzy składowymi funkcjonału $PRSS(f,\lambda)$.

W metodzie regresji grzbietowej można przyjąć:

$$RSS(f) = \sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{m} a_j x_j \right)^2, \quad J(f) = \sum_{j=1}^{m} a_j^2 . \tag{11}$$

Nałożenie kary (λ) na estymatory parametrów strukturalnych powoduje pomniejszenie ich wartości, aż do wyzerowania niektórych z nich [5]. W ten sposób z modelu liniowego zostają usunięte zmienne odpowiadające tym parametrom. Redukcja liczby zmiennych wprowadzanych do modelu implikuje uzyskanie prostszej funkcji regresji (z mniejszą liczbą składowych).

Regresja grzbietowa jest w literaturze przedstawiana w różnych kontekstach: jako próba rozwiązania problemu współliniowości zmiennych objaśniających, metoda szukająca kompromisu pomiędzy obciążeniem a wariancją estymatorów czy jako metoda doboru zmiennych do modelu. Są to niewątpliwie zalety tej metody. Interesującym wydaje się więc ocena jej zdolności predykcyjnych w kontekście innych skutecznych narzędzi wielowymiarowej analizy regresji, jakimi są metody nieparametryczne.

2. Analiza porównawcza

Głównym celem regresji jest predykcja. Ocena modelu regresyjnego nie powinna zatem być dokonywana na podstawie stopnia jego dopasowania do danych ze zbioru uczącego, tylko z wykorzystaniem miar pozwalających określić

zdolność predykcji tego modelu. Tak jak już wspomniano we wprowadzeniu, w niniejszym artykule pojęcie *zdolność predykcyjna modelu* będzie rozumiane jako ocena, na ile wartości teoretyczne $\hat{y}_i$ różnią się od wartości empirycznych y_i dla obserwacji spoza zbioru uczącego.

Do oceny jakości predykcji można wykorzystać zbiór testowy, jeśli jest dostępny, lub wydzielić ze zbioru danych *część uczącą* oraz *walidacyjną*. Trzecim alternatywnym podejściem jest zastosowanie metody *sprawdzania krzyżowego* (*cross validation*) [por. 9], [2], w której zbiór danych jest dzielony na b w przybliżeniu równolicznych oraz rozłącznych części uczących oraz testowych. W każdym z b kroków algorytmu tej metody, jedną (ale za każdym razem inną) część z otrzymanego podziału wykorzystuje się do testowania modelu, zbudowanego na pozostałych $b-1$ częściach zbioru danych. W ten sposób otrzymuje się b wartości miernika jakości predykcji modelu, którym najczęściej jest błąd średniokwadratowy MSE. Wartości te zostają następnie uśrednione, a otrzymana statystyka MSE_{CV} jest nieobciążonym estymatorem błędu średniokwadratowego [por. 9].

Zbiory danych wykorzystane w analizie

Analizę przeprowadzono na pięciu rzeczywistych zbiorach danych, standardowo wykorzystywanych do badania własności różnych metod regresji. Najważniejsze charakterystyki tych zbiorów zestawiono w tabeli 1.

Tabela 1

Charakterystyki zbiorów danych wykorzystywanych w analizie

Nazwa zbioru	Liczba obserwacji	Liczba zmiennych
Autompg	398	8
Boston	506	14
Clothing	400	13
Ozone	366	13
Star	5748	6

Zbiór danych *Autompg* pochodzi z repozytorium `StatLib` z uniwersytetu Carnegie Mellon[3], natomiast pozostałe zbiory danych są dostępne w bibliotekach `mlbench` oraz `Ecdat` programu statystycznego **R**.

[3] http://archive.ics.uci.edu/ml/datasets/Auto+MPG.

Nieparametryczne metody regresji wykorzystane w analizie

W badaniu porównywano model regresji grzbietowej z nieparametrycznymi modelami regresji zbudowanymi za pomocą:

- metody rzutowania PPR [4],
- metody zagregowanych drzew regresyjnych Breimana – RANDOM FORESTS [3],
- wielowymiarowej metody krzywych sklejanych POLYMARS [10],
- metody wykorzystującej sieci neuronowe (oznaczonej jako NNET) [por. 1].

Do budowy modeli regresji wykorzystano program statystyczny **R** z dodatkowymi bibliotekami. Większość badanych metod wymaga ustalenia wartości pewnych parametrów budowanego modelu regresji. Przeszukiwane zakresy parametrów dla poszczególnych metod to:

- w metodzie rzutowania PPR wartość parametru opisującego początkową liczbę funkcji składowych modelu przyjmowano na poziomie: 10, 15, 20, 25, zaś końcowa liczba tychże funkcji w modelu zmieniała się od 1 do 10,
- w metodzie zagregowanych drzew regresyjnych Breimana liczbę zmiennych losowanych przy każdym podziale ustalano na poziomie: $\sqrt{m}$, $\frac{m}{3}$, $2\sqrt{m}$ (m – liczba zmiennych), liczbę drzew równą 100 oraz 200, zaś minimalną liczbę obserwacji w liściu: 1, 5, 10,
- w modelach POLYMARS oraz regresji grzbietowej przyjęto domyślne wartości parametrów, zaproponowane przez funkcję realizującą tę metodę w programie statystycznym **R**,
- w modelach sieci neuronowych z jedną ukrytą warstwą, przyjmowano liczbę obserwacji w warstwie ukrytej zmieniającą się od 1 do ln(n) (gdzie n jest liczbą obserwacji).

Procedura badawcza

Do porównania, pod względem zdolności predykcyjnych, modelu regresji grzbietowej z modelami nieparametrycznymi, zostało zbudowanych wiele modeli. Były one tworzone dla różnych zestawów parametrów, dla każdej z metod. Jednak w ostatecznym zestawieniu daną metodę reprezentuje zawsze tylko jeden model – ten w którym wykorzystano optymalną kombinację parametrów. Zwieńczeniem tego etapu procedury badawczej jest zestawienie modeli, pod względem dokładności predykcji, ocenianej za pomocą estymatora punktowego, jakim jest błąd średniokwadratowy obliczony metodą sprawdzania krzyżowego (MSE_{CV}). Model będący najlepszym rozwiązaniem danego zadania regresji to ten o najmniejszej wartości błędu MSE_{CV}. Szczegółowo etapy procedury badawczej przedstawiono w tabeli 2.

Tabela 2

Kroki procedury badawczej – porównywanie zdolności predykcyjnych modeli za pomocą estymatora punktowego MSE_{CV}

Krok 1	Podziel zbiór uczący D na 10 równolicznych (w przybliżeniu) oraz rozłącznych części*
Krok 2	Wykonanie następujących czynności dla każdej z rozpatrywanych metod regresji: a) zbuduj wiele modeli regresji dla różnych wartości parametrów tej metody, b) oblicz metodą sprawdzania krzyżowego błąd MSE_{CV} dla wszystkich modeli otrzymanych w punkcie a), c) wybierz ten układ parametrów i odpowiadający mu model, dla którego uzyskałeś najmniejszy błąd MSE_{CV}; wybrany model jest reprezentantem danej metody do porównań
Krok 3	Stwórz zestawienie analizowanych modeli regresji, pod względem otrzymanych wartości błędów MSE_{CV}

* Możliwy jest podział zbioru danych na inną liczbę części, jednak Kohavi w pracy [9] zaleca stosowanie metody sprawdzania krzyżowego z parametrem $b \leq 10$.

2.4. Wyniki analizy

Po wyznaczeniu modelu optymalnego (najlepszego układu parametrów) dla każdej z metod i każdego z badanych zbiorów[4], zestawiono otrzymane wyniki w tabeli 3, według kryterium błędu średniokwadratowego obliczonego metodą sprawdzania krzyżowego. Najlepsze wyniki zaznaczono pogrubioną czcionką, natomiast najgorsze kursywą.

Tabela 3

Błędy średniokwadratowe MSE_{CV} obliczone dla modeli otrzymanych różnymi metodami regresji

	Autompg	*Boston*	*Clothing*	*Ozone*	*Star*
R. grzbietowa	*11,11*	*22,68*	$81654 \cdot 10^6$	*19,40*	*2 088,6*
PPR	7,62	10,31	$\mathbf{10525 \cdot 10^6}$	17,06	1 988,3
R.FOREST	**4,04**	**5,74**	$47579 \cdot 10^6$	**8,93**	**1 812,1**
POLYMARS	7,45	11,85	$\mathit{94507 \cdot 10^9}$	14,59	2 082,2
NNET	8,75	14,13	$68114 \cdot 10^6$	13,08	2 037,8

[4] Należy wspomnieć, że nie wszystkie metody regresji pozwalają na budowę modelu, gdy w zbiorze danych brakuje wartości niektórych zmiennych. W celu zapewnienia pełnej porównywalności otrzymanych modeli, ze zbiorów danych usunięto obserwacje z brakującymi wartościami.

Podsumowanie

Wyniki przeprowadzonych analiz pokazują, iż nie można wskazać metody regresji, która dawałaby najmniejsze błędy średniokwadratowe, niezależnie od rozważanego zbioru danych, choć najczęściej najlepsze wartości uzyskano dla metody zagregowanych drzew regresyjnych Breimana.

Wśród otrzymanych wyników, w czterech przypadkach na pięć, błędy MSE_{CV} modelu regresji grzbietowej są największe, a model ten zdecydowanie ustępujące modelom regresji nieparametrycznej pod względem zdolności predykcyjnych.

Literatura

[1] Bishop C., *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford 1995.

[2] Blum A., Kalai A., Langford J., *Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation*, „COLT" 1999, s. 203-208.

[3] Breiman L., *Random Forests*, „Machine Learning" 2001, Vol. 45, s. 5-32.

[4] Friedman J., Stuetzle W., *Projection Pursuit Regression*, „Journal of the American Statistical Association" 1981, Vol. 76, s. 817-823.

[5] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer Verlag, New York 2001.

[6] Hoerl A.E., Kennard R.W., *Ridge Regression: Applications to Nonorthogonal Problems*, „Technometrics" 1970, Vol. 12, s. 69-82.

[7] Hoerl A.E., Kennard R.W., *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, „Technometrics" 1970, Vol. 12, s. 55-67.

[8] Hothorn T., Leisch F., Zeileis A., Hornik K., *The Design and Analysis of Benchmark Experiments*, „Journal of Computational and Graphical Statistics" 2005, Vol. 14(3), s. 675-699.

[9] Kohavi R., *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, „IJCAI" 1995, s. 1137-1145.

[10] Kooperberg C., Bose S., Stone C., *Polychotomous Regression*, „Journal of the American Statistical Association" 1997, Vol. 92, s. 117-127.

[11] Maddala G.S., *Ekonometria*, WN PWN, Warszawa 2006.

[12] Trzęsiok J., *Porównanie nieparametrycznych modeli regresji pod względem zdolności predykcyjnych*, [w:] *Metody i modele analiz ilościowych w ekonomii i zarządzaniu* cz. 4, red. J. Mika, Wydawnictwo Uniwersytetu Ekonomicznego, Katowice 2012, s. 102-111.

[13] Welfe A., *Ekonometria*, PWE, Warszawa 2003.

COMPARING THE PERFORMANCE OF THE RIDGE REGRESSION WITH SOME NONPARAMETRIC REGRESSION MODELS

Summary

The paper presents a short description of ridge regression and comparing the performance of this regression with some nonparametric methods of regression. The analysis was conducted with the use of simulation procedures on benchmarking data sets.