

Joanna Trzęsiok
Akademia Ekonomiczna w Katowicach

ADAPTACYJNE METODY JĄDROWE W REGRESJI

1. Wstęp

Przedmiotem analizy jest zbiór obserwacji:

$$U = \{(x_i, y_i) : i = 1, 2, \dots, N\}.$$

Celem klasycznej analizy regresji jest znalezienie funkcji f , która opisywałaby zależność zmiennej Y od zmiennej objaśniającej X . Szuka się więc funkcji f , dla której:

$$Y = f(X) + \varepsilon, \quad (1)$$

gdzie ε jest składnikiem losowym.

W regresji jądrowej nie jest ważna analityczna postać f . Poszukuje się jedynie wartości funkcji f dla pojedynczych punktów x_0 :

$$y_0 = f(x_0) + \varepsilon. \quad (2)$$

Z tego względu metody te zalicza się do klasy metod adaptacyjnych.

2. Adaptacyjne metody regresji

Jak już wspomniano, w metodach adaptacyjnych często nie znamy analitycznej postaci funkcji f . Jeśli jest taka potrzeba, do modelu dołączane są funkcje składowe, których parametry są szacowane w kolejnych krokach. Istotną rolę odgrywa tu właśnie sposób estymacji współczynników funkcji f .

Metody adaptacyjne mają własności lepsze niż metody klasyczne. Modele otrzymywane za ich pomocą charakteryzują się często lepszym dopasowaniem do danych empirycznych oraz mniejszymi błędami prognoz. Nie zakładamy w tym przypadku ani znajomości rozkładów badanych cech (zmiennie objaśniające nie

muszą mieć rozkładów normalnych), ani analitycznych postaci związków między nimi. Pozwala to na dużo szersze zastosowanie tych metod.

3. Metody jądrowe w przestrzeni jednowymiarowej

Metody wykorzystywane w regresji jądrowej są zaliczane do metod **estymacji lokalnej**. Polegają one na szacowaniu wartości funkcji f dla pojedynczych punktów z osobna, a nie na szukaniu postaci funkcji f . Estymuje się więc wartości y_0 :

$$y_0 = \hat{f}(x_0). \quad (3)$$

Punkty x_0 są wybierane z przedziału zmienności zmiennej X . Metoda nie narzuca sposobu ich ustalania. Standardowo wykorzystywana procedura dzieli zakres zmienności zmiennej X na z góry ustaloną liczbę przedziałów o równej długości i przyjmuje otrzymane w ten sposób punkty podziału za punkty x_0 , w których estymowane będą wartości funkcji regresji.

W metodach regresji jądrowej za funkcję $\hat{f}$ przyjmujemy najczęściej wielomian. W najprostszym przypadku $\hat{f}$ jest funkcją stałą, a $\hat{f}(x_0)$ – średnią ważoną wartości y_i . Z takim podejściem mamy do czynienia m.in. w metodzie „ k -najbliższych sąsiadów” oraz w estymatorze Nadaraya-Watsona.

W metodach adaptacyjnych ważną rolę odgrywa również sposób estymacji współczynników funkcji $\hat{f}$. W regresji jądrowej parametry $\hat{f}$ szacujemy, opierając się jedynie na tych obserwacjach (x_i, y_i) , dla których x_i leży w otoczeniu punktu x_0 .

3.1. Metoda „ k -najbliższych sąsiadów”

W metodzie „ k -najbliższych sąsiadów” wartość y_0 jest obliczana jako średnia ważona wartości y_i , dla których współrzędne x_i leżą w sąsiedztwie x_0 :

$$y_0 = \hat{f}(x_0) = \frac{1}{k} \sum_{i=1}^N K_k(x_0, x_i) \cdot y_i, \quad (4)$$

gdzie $K_k(x_0, x_i)$ jest **funkcją jądrową** nadającą wagi wartościom y_i .

Jeśli przez $N_k(x_0)$ oznaczymy k -elementowe sąsiedztwo punktu x_0 , to funkcję jądrową $K_k(x_0, x_i)$ możemy zdefiniować następująco:

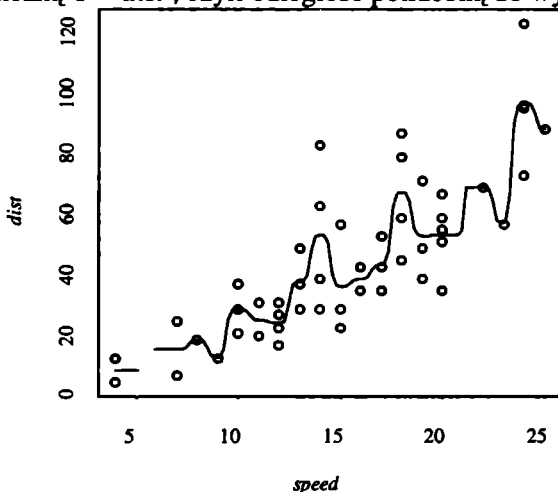
$$K_k(x_0, x_i) = \begin{cases} 1, & \text{dla } x_i \in N_k(x_0) \\ 0, & \text{dla } x_i \notin N_k(x_0) \end{cases} \quad (5)$$

Wartość y_0 jest więc obliczana na podstawie k wartości y_i . Wagi nadane y_i są stałe: jeden, gdy x_i należy do sąsiedztwa x_0 , zero – w przeciwnym razie.

Funkcja jądrowa $K_k(x_0, x_i)$ zależy od parametru k , czyli od liczby elementów sąsiedztwa $N_k(x_0)$. Jeśli k zostanie przyjęte *a priori*, to będziemy mieli do czynienia z metodą nieadaptacyjną. Jeżeli k wyznaczymy na podstawie zbioru uczącego, ale będzie ono stałe dla wszystkich punktów x_0 , to mamy do czynienia z adaptacyjną metodą globalną. W metodzie adaptacyjnej lokalnej k powinno być obliczone na podstawie zbioru uczącego dla każdego x_0 z osobna. To podejście jest jednak trudne i pracochłonne.

W celu zilustrowania metody „ k -najbliższych sąsiadów” przedstawiony został wykres dla zbioru danych *cars* (zob. [Ezekiel 1930]). Dane w tym zbiorze są scharakteryzowane przez dwie zmienne:

- zmienną objaśniającą $X = \text{speed}$, wyrażającą szybkość samochodu,
- zmienną zależną $Y = \text{dist}$, czyli odległość potrzebną do wyhamowania.



Rys.1. Wykres dla zbioru danych *cars* ilustrujący metodę „ k -najbliższych sąsiadów”.

Źródło: opracowanie własne z użyciem pakietu *R*.

Na rys. 1 znajdują się oszacowane na podstawie metody „ k -najbliższych sąsiadów” wartości y_0 , które połączone zostały krzywą.

3.2. Estymator Nadaraya-Watsona

Wartość y_0 oszacowana za pomocą estymatora Nadaraya-Watsona jest również średnią ważoną:

$$y_0 = \hat{f}(x_0) = \frac{\sum_{i=1}^N K_{\lambda}(x_0, x_i) \cdot y_i}{\sum_{i=1}^N K_{\lambda}(x_0, x_i)}, \quad (6)$$

gdzie $K_{\lambda}(x_0, x_i)$ jest funkcją jądrową zależną od odległości pomiędzy punktami x_0 i x_i oraz parametru λ :

$$K_{\lambda}(x_0, x_i) = D\left(\frac{|x_0 - x_i|}{\lambda}\right). \quad (7)$$

Oryginalnie w estymatorze Nadaraya-Watsona funkcja jądrowa oparta była na funkcji Epanechnikowa danej wzorem:

$$D(t) = \begin{cases} \frac{3}{4}(1-t^2), & \text{dla } |t| \leq 1 \\ 0, & \text{dla } |t| > 1 \end{cases}. \quad (8)$$

Wykorzystywane są również inne funkcje, wśród których najczęściej spotykamy

a) funkcję „*tri-cube*”:

$$D(t) = \begin{cases} (1-t^3)^3, & \text{dla } |t| \leq 1 \\ 0, & \text{dla } |t| > 1 \end{cases}, \quad (9)$$

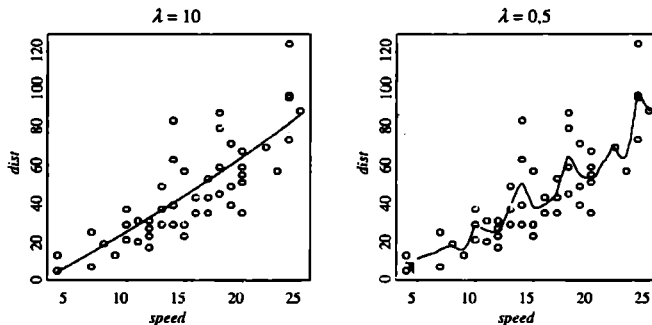
b) funkcję gęstości Gaussa:

$$D(t) = \Phi(t). \quad (10)$$

Ponieważ w metodzie tej funkcja jądrowa $K_{\lambda}(x_0, x_i)$ zależy od odległości pomiędzy x_0 i x_i , to na szacowaną wartość y_0 wpływają najbardziej te y_i , dla których x_i są najbliższe x_0 . Im x_i znajduje się dalej od x_0 , tym waga odpowiadającej mu współrzędnej y_i jest mniejsza. Jeśli zaś x_i znajduje się na zewnątrz otoczenia x_0 , to $K_{\lambda}(x_0, x_i)$ równe jest zeru.

Funkcja jądrowa $K_{\lambda}(x_0, x_i)$ zależy również od parametru λ , nazywanego często **parametrem szerokości okna** lub parametrem wygładzenia, wyrażającego promień otoczenia punktu x_0 . Przy wyborze parametru λ występuje kilka problemów:

- jeśli dobierzemy zbyt małe λ , to $\hat{f}(x_0)$ zależeć będzie jedynie od kilku wartości y_i i w efekcie otrzymamy funkcję zbyt dobrze dopasowaną do danych,
- jeśli ustalimy zbyt duże λ , to $\hat{f}(x_0)$ będziemy obliczać na podstawie prawie wszystkich y_i , a uzyskana funkcja nie będzie się w tej sytuacji różniła od liniowej funkcji regresji otrzymanej metodą najmniejszych kwadratów.



Rys. 2. Wykres funkcji regresji otrzymanej dla estymatora Nadaraya-Watsona dotyczący zbioru danych *cars* dla dwóch wybranych wartości parametru λ

Źródło: opracowanie własne z użyciem pakietu *R*.

Na rysunku 2 pokazano funkcje regresji, otrzymane za pomocą estymatora Nadaraya-Watsona dla zbyt dużego $\lambda = 10$ i zbyt małego $\lambda = 0,5$. Przedstawiono oszacowane wartości y_0 , które połączono krzywą.

3.3. Ogólny zapis problemu regresji jądrowej

Zarówno w metodzie „ k -najbliższych sąsiadów”, jak i w estymatorze Nadaraya-Watsona za funkcję $\hat{f}$ przyjmowaliśmy funkcję stałą. Ogólne podejście do problemu regresji jądrowej polega na szacowaniu wartości y_0 z użyciem wielomianu stopnia p przez rozwiązanie zadania minimalizacji lokalnego ryzyka empirycznego.

Lokalne ryzyko empiryczne (ważony błąd średniokwadratowy) dane jest wzorem:

$$R_{emp-local}(\beta(x_0)) = \frac{1}{N} \sum_{i=1}^N K_{\lambda}(x_0, x_i) \cdot [y_i - f(x_i, \beta(x_0))]^2, \quad (11)$$

gdzie $K_{\lambda}(x_0, x_i)$ jest, jak wcześniej, funkcją jądrową, a $f(x_i, \beta(x_0))$ – funkcją wykorzystywaną do estymacji lokalnej o parametrach wewnętrznych:

$$\beta(x_0) = (\beta_0(x_0), \beta_1(x_0), \dots, \beta_p(x_0)). \quad (12)$$

W adaptacyjnych metodach jądrowych w regresji estymator funkcji f jest najczęściej wielomianem stopnia p .

Zatem

$$y_0 = \hat{f}(x_0, \beta(x_0)) = \hat{\beta}_0(x_0) + \sum_{j=1}^p \hat{\beta}_j(x_0) \cdot x_0^j. \quad (13)$$

Wyznaczenie wartości y_0 sprowadza się więc do oszacowania wektora parametrów $\beta(x_0)$, które otrzymujemy jako rozwiązanie zadania minimalizacji lokalnego ryzyka empirycznego:

$$\min_{\beta(x_0)} \sum_{i=1}^N K_{\lambda}(x_0, x_i) \cdot \left[y_i - \beta_0(x_0) - \sum_{j=1}^p \beta_j(x_0) \cdot x_i^j \right]^2. \quad (14)$$

Nadanie wag kwadratowi różnic pomiędzy wartościami rzeczywistymi y_i , a teoretycznymi $\hat{f}(x_i, \beta(x_0))$ powoduje, że w estymowaniu parametrów $\beta(x_0)$ biorą udział tylko te obserwacje (x_i, y_i) , dla których x_i leży w otoczeniu punktu x_0 , określonym przez funkcję jądrową $K_{\lambda}(x_0, x_i)$.

Minimalizacja lokalnego ryzyka empirycznego prowadzi do wyznaczenia estymatora parametrów $\hat{\beta}(x_0)$ (zob. [Wand, Jones 1995]):

$$\hat{\beta}(x_0) = (\mathbf{B}^T \mathbf{W}(x_0) \mathbf{B})^{-1} \mathbf{B}^T \mathbf{W}(x_0) \mathbf{y}, \quad (15)$$

gdzie:

$$\mathbf{B} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^p \end{bmatrix}, \mathbf{W} = \begin{bmatrix} K_\lambda(x_0, x_1) & 0 & \dots & 0 \\ 0 & K_\lambda(x_0, x_2) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & K_\lambda(x_0, x_N) \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

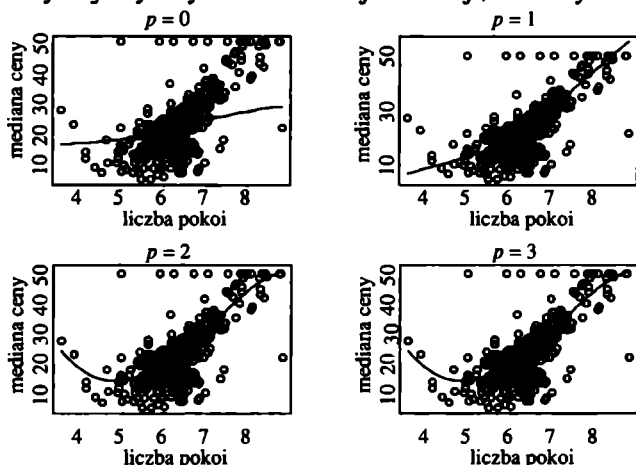
W związku z tym otrzymujemy oszacowanie wartości y_0 :

$$\hat{f}(x_0) = \mathbf{e}_1^T (\mathbf{B}^T \mathbf{W}(x_0) \mathbf{B})^{-1} \mathbf{B}^T \mathbf{W}(x_0) \mathbf{y}. \quad (16)$$

Łatwo pokazać, że dla $p = 0$, czyli stałej funkcji $\hat{f}$ rozwiązaniem zadania (14) – minimalizacji lokalnego ryzyka empirycznego – jest estymator Nadaraya-Watsona dany wzorem (6).

Estymowana lokalnie wartość y_0 zależy od stopnia wielomianu $\hat{f}$ oraz parametru λ . Wybór zarówno p , czyli stopnia wielomianu, jak i parametru λ jest kompromisem pomiędzy obciążeniem a wariancją estymatora.

- Jeśli λ jest zbyt małe, to $\hat{f}(x_0)$ obliczamy na podstawie kilku wartości y_i . Wartości teoretyczne są wtedy bliskie wartościom rzeczywistym, dlatego obciążenie estymatora jest niskie. Jednak ceną jest wysoka wariancja.
- Jeśli λ jest zbyt duże, to $\hat{f}(x_0)$ obliczamy na podstawie wielu wartości y_i i w efekcie otrzymujemy estymator o niskiej wariancji, ale o wysokim obciążeniu.



Rys. 3. Zastosowanie wielomianów różnych stopni w estymacji lokalnej
Źródło: opracowanie własne.

Zastosowania wielomianów różnych stopni do estymacji lokalnej zaprezentowane będzie na przykładzie zbioru danych *Boston* (zob. [Harrison, Rubinfeld 1978]). Jest to zbiór znany i wykorzystywany do sprawdzania jakości modeli regresyjnych. Zgromadzone dane są charakteryzowane przez kilkanaście zmiennych,

jednak do analizy wybrano liczbę pokoi jako zmienną objaśniającą X . Zmienną zależną Y jest mediana wartości domu w tys. dol.

Na rysunku 3 pokazano funkcję regresji otrzymaną z wykorzystaniem metod jądrowych, zaimplementowanych w pakiecie statystycznym R dla wielomianów stopnia $p = 0, 1, 2$ i 3 .

4. Uogólnienie metod jądrowych na przestrzeń wielowymiarową

Przedmiotem analizy jest teraz zbiór obserwacji:

$$U = \{(\mathbf{x}_i, y_i) = (x_1^i, x_2^i, \dots, x_n^i, y_i) : i = 1, 2, \dots, N\}. \quad (17)$$

Każdy element tego zbioru jest charakteryzowany przez wektor $n + 1$ zmiennych:

$$(X_1, X_2, \dots, X_n, Y).$$

Estymowaną lokalnie wartość $y_0 = \hat{f}(\mathbf{x}_0)$ otrzymujemy jako rozwiązanie zadania minimalizacji lokalnego ryzyka empirycznego:

$$\min_{\beta(\mathbf{x}_0)} \sum_{i=1}^N K_{\lambda}(\mathbf{x}_0, \mathbf{x}_i) \cdot [y_i - f(\mathbf{x}_i, \beta(\mathbf{x}_0))]^2, \quad (18)$$

gdzie $f: \mathbf{R}^n \rightarrow \mathbf{R}$ jest funkcją o wewnętrznych parametrach:

$$\beta(\mathbf{x}_0) = (\beta_0(\mathbf{x}_0), \beta_1(\mathbf{x}_0), \dots, \beta_m(\mathbf{x}_0)).$$

Funkcja jądrowa jest zależna od odległości pomiędzy punktami $\mathbf{x}_0$ i $\mathbf{x}_i$ oraz parametru λ :

$$K_{\lambda}(\mathbf{x}_0, \mathbf{x}_i) = D\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_i\|}{\lambda}\right), \quad (19)$$

gdzie D jest np. funkcją Epanechnikowa.

5. Przykład zastosowania regresji jądrowej

W celu zilustrowania adaptacyjnych metod jądrowych w regresji ponownie przeprowadzono obliczenia na zbiorze *Boston* za pomocą programu statystycznego R .

Obserwacje przedstawione w zbiorze *Boston* zostały zebrane i opublikowane w 1978 r. przez D. Harrisona i D. Rubinfeld, badaczy, którzy zajmowali się wykrywaniem zależności pomiędzy cenami nieruchomości w Bostonie a jakością życia.

Lokalne szacowanie wartości y_0 za pomocą metod jądrowych często daje wysokie obciążenie estymatora na końcach przedziału zmienności zmiennej X . W regresji jądrowej w przestrzeni wielowymiarowej problem ten znacznie się nasila. W związku z tym opisywane metody są wykorzystywane w praktyce jedynie do przestrzeni dwu- lub trójwymiarowej. Dlatego też do analizy wykorzystano tylko dwie zmienne objaśniające:

X_1 – liczbę pokoi,

X_2 – procent ludności o niskim statusie społecznym.

Zmienną zależną Y jest mediana wartości domu w tys. dol.

Wykorzystany model lokalnej funkcji regresji ma postać:

$$y_0 = \hat{f}(x_1^0, x_2^0),$$

gdzie $\hat{f}$ jest wielomianem pierwszego stopnia. Wartości teoretyczne y_0 wyznaczone za pomocą metod jądrowych zaimplementowanych w programie *R*. Dopasowanie modelu obliczone za pomocą współczynnika R^2 wynosi 0,768.

Dla porównania obliczono również R^2 dla modelu liniowego, modelu zbudowanego metodą rzutowania (ang. *projection pursuit regression*) oraz otrzymanego metodą MARS. Wyniki przedstawia tab. 1.

Tabela 1. Dopasowanie różnych funkcji regresji, obliczone za pomocą współczynnika R^2

Model	Funkcja liniowa	MARS	Metoda rzutowania	Regresja jądrowa
R^2	0,637	0,793	0,796	0,768

Źródło: opracowanie własne.

Współczynnik dopasowania modelu uzyskanego za pomocą metod jądrowych nie jest wprawdzie najwyższy, ale bliski R^2 dla modelu zbudowanego metodą rzutowania, dającą najczęściej najlepsze wyniki.

Literatura

- Cherkassky V., Mulier F., *Learning from Data – Concepts, Theory, and Methods*, John Wiley & Sons, Inc., New York 1998.
- Ezekiel M., *Methods of Correlation Analysis*, John Wiley & Sons, Inc., 1930.
- Harrison D., Rubinfeld D.L., *Hedonic Prices and the Demand for Clean Air*, „Journal of Environmental Economics and Management” 1978, 8.
- Hastie T., Tibshirani R., Friedman J.H., *The Elements of Statistical Learning*, Springer-Verlag, New York 2001.
- Ruppert D., Sheather S.J., Wand M.P., *An Effective Bandwidth Selector for Local Least Squares Regression*, „Journal of the American Statistical Association” 1995, 90, s. 1257-1270.
- Venables W.N., Ripley B.D., *Modern Applied Statistics with S-PLUS*, Springer-Verlag New York 1997.
- Wand M.P., Jones M.C., *Kernel Smoothing*, Chapman and Hall, London 1995.

ADAPTIVE KERNEL METHODS OF REGRESSION

Summary

In this article kernel methods for regression are presented. These methods are based on theory of local estimation when one needs to estimate an unknown function only at a single point. The methods don't need the analytic form of the model, so they are the adaptive methods. An example is given for illustration.