METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XII/2, 2011, str. 202–209

ZANURZANIE W REGRESJI LINIOWEJ

Małgorzata Kobylińska
Katedra Metod Ilościowych
Uniwersytet Warmińsko-Mazurski w Olsztynie
e-mail: angosiak@poczta.onet.pl

Streszczenie: Wprowadzone przez Tukey'a [Tukey 1975] pojecie zanurzania obserwacji w próbach wielowymiarowych stało się narzędziem służącym analizie danych. Dzięki wykorzystaniu miary zanurzania obserwacji w próbie przezwycięża się trudności związane z porządkowaniem obserwacji wielowymiarowych. Pojęcie zanurzania danych było intensywnie rozwijane przez wielu badaczy z punktu jego przydatności do opisu statystycznego danych jedno i wielowymiarowych W literaturze przedmiotu spotkać można różne kryteria oraz metody wyznaczania miary zanurzania obserwacji w próbie. W pracy podano określenie zanurzenia obserwacji w próbie oraz pojęcia z nim związane. Przedstawiono wykorzystanie zanurzania w regresji liniowej dla przypadku dwuwymiarowego.

Słowa kluczowe: zanurzanie obserwacji w próbie, funkcja regresji liniowej

WPROWADZENIE

Praca Tukey'a [Tukey 1975] stała się inspiracją do wprowadzenia wielu pojęć związanych z analizą eksploratywną danych liczbowych. Jednym z nich jest zanurzanie danych, będące relatywną miarą obserwacji w danym zbiorze danych.

W 1998 roku Rousseeuw i Huber wprowadzili pojecie zanurzania funkcji regresji liniowej w zbiorze dwuwymiarowym. Wartość zanurzania regresji jest liczbą całkowitą należącą do przedziału od zera do n i może być wykorzystywana do badania dopasowania równania regresji liniowej do danych empirycznych, przy czym wyższe wartości zanurzania odpowiadają funkcjom lepiej dopasowanym do tych danych.

W pracy przedstawione zostanie określenie zanurzania obserwacji w próbie oraz pojęcia z nim związane. Zanurzanie w próbie dwuwymiarowej rozważane będzie, jako pewna własność dopasowania regresji liniowej do danych

empirycznych dla przypadku dwuwymiarowego. Metoda wyznaczania zanurzania regresji liniowej zilustrowana będzie na przykładzie liczbowym. Zaprezentowany zostanie algorytm umożliwiający wyznaczenie zanurzania funkcji regresji w zbiorze dwuwymiarowym.

OKREŚLENIE ZANURZANIA OBSERWACJI W PRÓBIE

Niech $P_n^p = \{x_1, x_2, ..., x_n\}$ będzie układem obserwowalnych wektorów wyrażających próbę *p*-wymiarową o liczebności *n* pochodzącą z pewnego *p*-wymiarowego rozkładu określonego dystrybuantą F_p oraz niech $\theta \in R^p$ będzie pewnym punktem z przestrzeni rzeczywistej R^p. W szczególności może należeć on do układu punktów z próby P_n^p. Wówczas każdy punkt x_i jest rozpatrywany jako p-wymiarowy wektor kolumnowy $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]^T$, gdzie x_{ij} jest wartością *j*-tej zmiennej (czyli zmiennej X_j) zaobserwowaną dla *i*-tego obiektu przy $i = 1,2,...,n$ oraz $j = 1,2,...,p$. Jeżeli nie więcej niż *p* obserwacji należy do jakiejkolwiek (*p-1*)-wymiarowej podprzestrzeni, to próbę P_n^p nazywany zbiorem punktów ogólnie pozytywnym, według nazewnictwa wprowadzonego przez Donoho i Gasko [Donoho i Gasko 1992].

Dla określenia zanurzania obserwacji w próbie należy zauważyć, iż wektory próby $x_i \in P_n^p$ mogą być uporządkowane, gdy zostaną one przekształcone do wielkości skalarnych $u^T x_i$ dla zadanego wektora $u \in R^p$. Wówczas ciąg wektorów $x_1, x_2, ..., x_n \in P_n^p$ przechodzi w ciąg skalarów $u^T x_1, u^T x_2, ..., u^T x_n \in R$, który może być uporządkowany monotonicznie. Wektor $\theta \in R^p$ sprowadza się do skalara $u^T \theta \in R$. Ważną kwestią jest zlokalizowanie $u^T \theta$ w ciągu $u^T x_1, u^T x_2, ..., u^T x_n$. Zanurzanie wektora θ w próbie P_n^p sprowadza się do badania odległości $u^T \theta$ od końców próby uporządkowanej niemalejąco $\{u^T x_{(i)} : i = 1,2,...,n\}$, gdzie $u^T x_{(1)} \le u^T x_{(2)} \le ... \le u^T x_{(n)}$. Na podanej koncepcji określona jest definicja zanurzania Tukey'a obserwacji w próbie *p*-wymiarowej.

Definicja 1. Zanurzaniem punktu θ w próbie P_n^p nazywamy funkcję próby $zan_p(\theta, P_n^p)$ o następującej postaci

$$zan_p(\theta, P_n^p) = \min_{\|u\|=1} \#\left\{ i : u^T \theta \geq u^T x_i \right\}, \tag{1}$$

gdzie u jest wektorem kolumnowym z przestrzeni R^p o długości 1, czyli o normie euklidesowej $\|u\| = 1$, natomiast $\#\{\cdot\}$ oznacza liczebność rozważanego zbioru.

Przez $zan_p(\theta, P_n^p)$ rozumiemy próbkową wersję zanurzania obserwacji w próbie P_n^p pochodzącej z rozkładu określonego dystrybuantą F_p.

Intuicyjnie zanurzanie punktu θ w próbie P_n^p wyraża najmniejsza liczba punktów z tej próby, położonych po jednej stronie wektora θ. Pozwala to na dokonanie porządkowania elementów próby P_n^p w ciąg wektorów $x_{(1)}, x_{(2)}, ..., x_{(n)}$ taki, że spełniony będzie ciąg nierówności $zan_p(x_{(1)} : P_n^p) \geq zan_p(x_{(2)} : P_n^p) \geq \ ... \ \geq zan_p(x_{(n)} : P_n^p)$, czyli ciąg wartości nierosnących. Obserwacje, którym odpowiadają wyższe wartości zanurzania zlokalizowane są bardziej centralnie w badanej próbie, te którym odpowiadają najniższe wartości tej miary są znacznie oddalone od centralnego skupienia próby.

W literaturze przedmiotu spotkać można różne kryteria wyznaczania zanurzania obserwacji w próbie m. in. kryterium Mahalanobisa, kryterium Oja lub kryterium Barnetta [patrz np. Liu i in. 1999 a także Kobylińska 2003]. Należy zaznaczyć, że zagadnienia związane z wyznaczaniem zanurzania obserwacji w próbie są bezpośrednio związane z pojęciem konturów zanurzania, które stanowią ich graficzne uzupełnienie. Określenie konturów zanurzania oraz metoda ich wyznaczania przedstawione zostały m. in w pracy Ruts i Rousseeuw [1996].

WYKORZYSTANIE ZANURZANIA W REGRESJI LINIOWEJ

Analiza regresji zajmuje się opisywaniem zależności pomiędzy zmiennymi. Konstruowane są modele, które ilościowo opisują związki pomiędzy tymi zmiennymi. Pozwalają one na analizę struktury zależności, znaczenia czynnika losowego oraz umożliwią prognozowanie. Równaniem regresji nazywamy równanie opisujące związek pomiędzy zmiennymi z uwzględnieniem występowania składnika losowego.

Rozpatrzymy zbiór dwuwymiarowy ze względu na dwie zmienne X i Y, o których wiadomo, że zmienna X wywiera trwały wpływ na zmienną Y. Szacujemy równanie regresji liniowej dla zaobserwowanych w próbie wartości badanych zmiennych. Oszacowanie funkcji regresji liniowej Y względem X przedstawia równanie

$$\hat{y}_i = ax_i + b,$$

gdzie:

$\hat{y}_i$ - teoretyczne wartości funkcji regresji odpowiadające danemu poziomowi realizacji zmiennej X,

x_i - zaobserwowane w próbie realizacje zmiennej objaśniającej X,

a, b - oceny parametrów funkcji regresji Y od X, przy czym a jest oceną współczynnika regresji liniowej, b - oceną wyrazu wolnego.

Reszty modelu równania regresji będące różnicą pomiędzy rzeczywistymi wartościami y_i i odpowiadającymi im wartościami teoretycznymi $\hat{y}_i$ wyznaczane są według wzoru

$$e_i = y_i - \hat{y}_i .$$

Niech $P_n^2 = \{(x_i, y_i); i = 1,2,...,n\} \subset R^2$ będzie próbą dwuwymiarową o liczebności n oraz niech $\hat{y}_i = ax_i + b$ będzie równaniem regresji dwóch zmiennych, oszacowanym dla zaobserwowanych w próbie P_n^2 realizacji zmiennych X i Y.

Estymacja parametrów równania regresji jest postępowaniem mającym na celu znalezienie ocen parametrów strukturalnych na podstawie danych z próby. Do tego celu wykorzystuje się zazwyczaj Klasyczną Metodę Najmniejszych Kwadratów, która pozwala na znalezienie takich ocen tych parametrów, że suma kwadratów odchyleń pomiędzy rzeczywistymi i teoretycznymi wartościami zmiennej objaśnianej jest najmniejsza. Estymacja parametrów liniowej funkcji regresji polega na znalezieniu takich wartości dla ich ocen, żeby model regresji był jak najlepiej dopasowany do danych empirycznych. W celu określenia jakości dopasowania funkcji regresji do tych danych wykorzystać można na przykład takie miary jak: wariancję resztową, odchylenie standardowe reszt, współczynnik zmienności losowej lub współczynnik determinacji. Do oceny jakości ocen parametrów strukturalnych służą odpowiednie testy statystyczne. W analizie liniowej funkcji regresji reszty tworzą rozkład empiryczny, którego rozpoznanie jest możliwe po oszacowaniu parametrów strukturalnych tej funkcji. Poprawnie skonstruowany model funkcji regresji, poza wysokim stopniem odzwierciedlania wartości empirycznych, powinien również charakteryzować się pewnymi pożądanymi własnościami rozkładu reszt (np. stałością wariancji lub zgodnością z rozkładem normalnym składnika losowego). Badanie losowości reszt ma na celu weryfikację hipotezy o trafności doboru postaci analitycznej równania regresji. Należy zatem sprawdzić, czy funkcja regresji liniowej trafnie opisuje zależność pomiędzy zmiennymi X i Y. Idea testu serii, który może być wykorzystywany w tym celu, uwzględnia znaki reszt równania regresji. Jeżeli kolejno według rosnących wartości zmiennej objaśniającej następują dość długie ciągi reszt, złożone wyłącznie z wartości o tych samych znakach (serie reszt), oznacza to, że reszty mają charakter nielosowy. Mała liczba serii odpowiada

sytuacji, w której punkty empiryczne nie układają się w sposób losowy poniżej i powyżej prostej regresji [patrz np. Goryl i in. 2007 lub Luszniewicz i Słaby 2008].

Jeżeli ocena dopasowania funkcji regresji do danych empirycznych budzi wątpliwości, należy zbadać przyczyny tego stanu rzeczy. Powodem może być na przykład występowanie obserwacji nietypowych w danym zbiorze danych lub niewłaściwa postać analityczna równania regresji.

W pracy zaproponowana zostanie metoda wyznaczania zanurzania równania regresji w próbie dwuwymiarowej, która może być wykorzystywana w celu zbadania poprawności doboru liniowej funkcji regresji do danych empirycznych.

Określenie funkcji, która jest niedopasowana do danych empirycznych zbioru dwuwymiarowego podaje definicja.

Definicja 2. Funkcję liniową $y = ax + b$ nazywamy niedopasowaną do danych empirycznych zbioru dwuwymiarowego, jeżeli dla każdego x_i istnieje liczba rzeczywista $\upsilon_y = \upsilon$ taka, że dla każdego $x_i \neq \upsilon$ zachodzi

$$e_i < 0 \text{ dla każdego } x_i < \upsilon \text{ i } e_i > 0 \text{ dla każdego } x_i > \upsilon$$

lub

$$e_i > 0 \text{ dla każdego } x_i < \upsilon \text{ i } e_i < 0 \text{ dla każdego } x_i > \upsilon .$$

Definicja 3. Zanurzaniem funkcji regresji $rzan(\hat{y}_i, P_n^2)$ w zbiorze dwuwymiarowym P_n^2 nazywamy najmniejsza liczbę obserwacji tego zbioru, które należy z niego usunąć, żeby prosta regresji stała się niedopasowana do danych empirycznych.

Zgodnie z definicją 2 zanurzanie regresji liniowej nie uwzględnia wartości reszty tylko jej znak.

Dla zbioru P_n^2 maksymalna wartość zanurzania funkcji regresji spełnia nierówność

$$\max rzan(\hat{y}_i, P_n^2) \geq \left\lceil \frac{n}{3} \right\rceil, \tag{2}$$

gdzie $\lceil A \rceil$ jest częścią całkowita liczby A. Wartość zanurzania będzie równa n, jeżeli wszystkie punkty P_n^2 będą leżały na tej prostej. Można przyjąć, że jeżeli wartość zanurzania funkcji regresji liniowej w P_n^2 będzie większa lub równa $\left\lceil \frac{n}{3} \right\rceil$, to funkcja liniowa trafnie opisuje zależność pomiędzy badanymi zmiennymi. Jeżeli zbiór P_n^2 spełnia warunek zbioru ogólnie pozytywnego, to maksymalna wartość

zanurzania jest nie większa od $\left\lceil \frac{n+2}{2} \right\rceil$. Własności dotyczące zanurzania funkcji regresji w zbiorach dwuwymiarowych i wielowymiarowych zostały szczegółowo omówione miedzy innymi w pracy Rousseeuw i Hubert 1998.

Algorytm obliczania zanurzania funkcji regresji liniowej w zbiorze dwuwymiarowym obejmuje następujące kroki [Huber i Rousseeuw 1998]:

Krok 1. Dokonujemy porządkowania obserwacji zbioru dwuwymiarowego P_n^2 względem wartości x_i takich, że $x_i \le x_2 \le ... \le x_{n-1} \le x_n$, dla każdego $i = 1,2,...,n.$

Krok 2. Wyznaczamy liczebności zbiorów $L^+(\upsilon)$, $L^-(\upsilon)$, $R^+(\upsilon)$, $R^-(\upsilon)$ jako

$$L^+(\upsilon) = \#\{i; x_i \le \upsilon \; i \; r_i \ge 0\},\; L^-(\upsilon) = \#\{i; x_i \le \upsilon \; i \; r_i \le 0\}$$
$$R^+(\upsilon) = \#\{i; x_i > \upsilon \; i \; r_i \ge 0\},\; R^-(\upsilon) = \#\{i; x_i > \upsilon \; i \; r_i \le 0\}$$

gdzie υ jest pewną liczbą rzeczywistą,

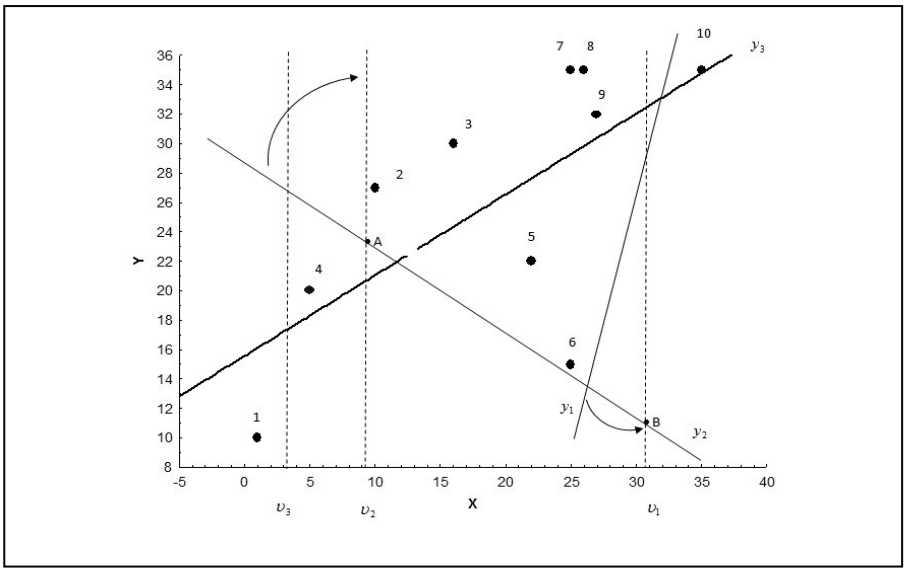
Krok 3. Obliczamy wartość zanurzania regresji liniowej w zbiorze P_n^2 zgodnie z wzorem

$$rzan(\hat{y}_i, P_n^2) = \min_{\upsilon}\left(\min\left\{L^+(\upsilon) + R^-(\upsilon), L^-(\upsilon) + R^+(\upsilon)\right\}\right).$$

Na wykresie korelacyjnym (Rys.1) umieszczono dwie funkcje liniowe y_1, y_2 oraz prostą regresji y_3 oszacowaną dla danych zbioru dwuwymiarowego P_{10}^2.

Zauważyć można, że funkcje y_1 oraz y_2 są niedopasowane do danych empirycznych P_{10}^2. Istnieją liczby rzeczywiste $x = \upsilon_1$ i $x = \upsilon_2$ takie, że w wyniku obrotu tych funkcji do pozycji pionowej względem osi X, odpowiednio dookoła punktów A i B, proste te nie przejdą przez żaden punkt P_{10}^2. Spostrzec można, że poprzez usuniecie obserwacji 5, 6 i 10 funkcja regresji y_3 stanie się funkcją niedopasowaną do danych empirycznych zbioru, czyli $rzan(y_3, P_{10}^2) = 3.$ Zgodnie z wzorem (2), maksymalna wartość zanurzania funkcji regresji w zbiorze P_{10}^2 jest większa lub równa 3, czyli postać liniowa funkcji regresji została dobrana poprawnie Badając losowość reszt testem serii otrzymujemy liczbę serii równą 4, wartości krytyczne odczytane z tablic rozkładu serii dla $\alpha = 0{,}05$ wynoszą odpowiednio 2 i 7. Nie ma więc podstaw do odrzucenia hipotezy głoszącej, że reszty modelu funkcji regresji liniowej mają charakter losowy.

Rysunek 1. Wykres korelacyjny P_{10}^2 z dwiema funkcja niedopasowanymi do tego zbioru oraz funkcją regresji oszacowaną dla obserwacji P_{10}^2



Źródło: opracowanie własne na podstawie [Hubert i Rousseeuw, 1998]

PODSUMOWANIE

W pracy przedstawiono metodę wyznaczania zanurzania regresji liniowej w próbie dwuwymiarowej. Może być ona wykorzystywana w celu zbadania dopasowania tej funkcji do danych empirycznych, przy czym wyższe wartości zanurzania świadczą o lepszym jej dopasowaniu. Przyporządkowanie danym funkcjom liniowym odpowiadających im wartości zanurzania pozwala na dokonanie rangowania tych funkcji względem ich dopasowania do rozważanych danych. Linia niedopasowana do danych empirycznych nigdy nie przechodzi przez żadną obserwację danego zbioru, ponieważ zgodnie z definicją 2 wszystkie składniki resztowe są ściśle dodatnie lub ściśle ujemne. Funkcja liniowa przechodząca przez k obserwacji zbioru dwuwymiarowego ma wartość zanurzania co najmniej k.

Zanurzanie regresji liniowej, jak wynika z definicji 2, uwzględnia tylko znak reszty. Można oczekiwać, że wysoka wartość zanurzania może wystąpić przy niskiej wartości współczynnika determinacji. Gorsze dopasowanie funkcji regresji do danych empirycznych może być konsekwencją wykorzystania do budowy tej funkcji danych zawierających obserwacje nietypowe. Obserwacje te zmieniają i wypaczają charakter zależności między badanymi zmiennymi, dlatego ważnym

zagadnieniem badawczym jest ich wykrycie i eliminacja we wstępnej analizie danych [patrz.np. Pawełek i Zeliaś 1996].

Zakres tego działu statystyki jest dość szeroki, w związku z tym trudno było omówić więcej problemów natury metodologicznej jak i empirycznej. Zasadniczą kwestią jest opracowanie efektywnego algorytmu numerycznego wyznaczania zanurzania funkcji regresji w zbiorach dwumiarowych i wielowymiarowych.

BIBLIOGRAFIA:

Donoho D.L., Gasko M. (1992) Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness, The Annals of Statistics, 20, 1803-1827.

Goryl A., Jędrzejczyk Z., Osiewalski J., Walkosz A. (2007) Wprowadzenie do ekonometrii w przykładach i zadaniach, PWN, Warszawa.

Hubert M., Rousseeuw P. J. (1998) The Catline for deep regression, Journal of Multivariate Analysis 66, 270-296.

Kobylińska M. (2006) Comparison of selected criteria for determination of the measure of depth of an observation in a two-dimensional sample, Acta Universitatis Lodziensis, Folia Oeconomica, 196.

Liu R.Y, Parelius J.M., Singh K. (1999) Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference, The Annals of Statistics, 27, 783-858.

Luszniewicz A., Słaby T., (2008) Statystyka z pakietem komputerowym STATISTICA PL Teoria i zastosowania, Wydawnictwo C.H.BECK.

Pawełek B., Zeliaś A. (1996) Obserwacje nietypowe w badaniach ekonometrycznych, Badania operacyjne i decyzje, nr 2, 59-86.

Rousseeuw R.J., Hubert M. (1998) Regression Depth, Journal of the American Statistical Association, 94, 388-402.

Rousseeuw P.J., Ruts I. (1996) Bivariate Location Depth, Applied Statistics , 45, 516-526.

Tukey J.W. (1975) Mathematics and the Picturing of Data, Proceedings of the International Congress of Mathematicians, 523-531.

DEPTH IN LINEAR REGRESSION

Abstract: The notion of observation depth in multidimensional samples introduced by Tukey [Tukey 1975] has become a new tool for data analysis. Applying the measure of observation depth in the sample the difficulties related to organisation of multidimensional observation are overcome. The notion of data depth has been developed extensively by many researchers from the perspective of its suitability for statistical description of single-dimensional and multidimensional data. In the subject literature different criteria and methods for determining the observation depth in the sample can be found. The paper presents the definition of the observation depth in the sample and the related notions. The application of depth in linear regression for a two-dimensional case was presented.

Key words: depth of the observation in a sample, linear regression function