METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XX/2, 2019, s. 83 – 92

Received: 21.01.2019
Accepted: 15.06.2019

METODA ZANURZANIA REGRESJI W PRZYPADKU WYSTĘPOWANIA OBSERWACJI NIETYPOWYCH

Małgorzata Kobylińska https://orcid.org/0000-0001-9674-5418
Wydział Nauk Ekonomicznych
Uniwersytet Warmińsko-Mazurski w Olsztynie
e-mail: gosiak@uwm.edu.pl

Streszczenie: W pracy przedstawiono wykorzystanie metody maksymalnego zanurzania regresji do szacowania parametrów strukturalnych funkcji regresji liniowej. Dla zbiorów dwuwymiarowych, zawierajacych obserwacje nietypowe, oszacowane zostały funkcje regresji wykorzystując klasyczną metodę najmniejszych kwadratów oraz metodę opartą na koncepcji zanurzania obserwacji w próbie. Zauważyć można w jaki sposób obserwacje nietypowe wpływają na oszacowane modele.

Słowa kluczowe: funkcja regrsji liniowej, klasyczna metoda najmniejszych kwadratów, zanurzanie obserwacji w próbie, zanurzanie regresji liniowej

JEL classification: C18, C40

WSTĘP

W literaturze coraz częściej poruszane są zagadnienia dotyczące statystyk odpornych oraz obserwacji nietypowych. Odporność jest rozumiana jako brak wrażliwości odnoszącej się do niespełnienia założeń dotyczących zgodności z rozkładem normalnym lub odporności na występowanie w danym zbiorze danych obserwacji nietypowych [Hampel 2000].

Założenie jednorodności zbioru danych jest jednym z podstawowych założeń dotyczących stosowania metod regresji. Zbiór punktów przestrzeni dwuwymiarowej nazywamy jednorodnym w sensie regresji liniowej, jeżeli tworzy w tej przestrzeni figurę o takim kształcie, że można ją aproksymować wykorzystując regresję liniową [Jajuga 1993].

Obserwacje nietypowe mogą zmieniać charakter zależności pomiędzy zmiennymi, co ma istotne znaczenie w badaniu zjawisk ekonomicznych. Z innej

https://doi.org/10.22630/MIBE.2019.20.2.9

strony mogą dostarczać ważnych informacji dotyczących zmian zachodzących w przypadku tych zjawisk. Jeżeli w danym w zbiorze występują obserwacje nietypowe (odstające) można zdecydować się na ich wyeliminowanie lub zastosowanie odpowiednich metod analizy statystycznej [Barnett, Lewis 1978].

Klasyczne metody statystycznej analizy wielowymiarowej nie są odporne na występowanie obserwacji nietypowych. Wyniki analizy, uzyskane z wykorzystaniem tych metod, mogą prowadzić do błędnych wniosków lub mogą być konsekwencją konstrukcji modelu funkcji regresji, który nieprawidłowo będzie opisywał badane zjawiska. Na szczególną uwagę zasługuje problem odporności w przypadku modeli regresji. Metodę regresji można uznać za odporną, jeżeli model regresji nie jest wrażliwy na występowanie obserwacji nietypowych. Wykazuje on tendencję, która jest reprezentowana przez większość obserwacji zbioru danych. Zagadnienia dotyczące regresji odpornej opisane zostały między innymi w pracach takich autorów jak Andersen [2008], Kosiorowski [2012] lub Rousseeuw i Leroy [1987].

Metody analizy regresji, oparte na koncepcji zanurzania obserwacji w próbie, należą do metod odpornych. W pracy przedstawiono wykorzystanie estymatora maksymalnego zanurzania funkcji regresji do szacowania parametrów modelu regresji dla danych dwuwymiarowych [Hubert, Rousseeuw 1999]. Modele regresji zostały oszacowane dla zbiorów liczbowych zawierających obserwacje nietypowe. W celu zilustrowania użyteczności metod opartych na zanurzaniu obserwacji w próbie, w przypadku występowania obserwacji odstających, oszacowane zostały również funkcje regresji wykorzystując klasyczną metodę najmniejszych kwadratów.

METODA BADAWCZA

W badaniu zjawisk ekonomicznych często należy określić powiązania pomiędzy zmienna zależną a zmiennymi niezależnymi. W tym celu wykorzystać można analizę regresji. Jako narzędzie badawcze pozwala ona zrozumieć i opisać analizowane zjawiska, określić siłę, kształt i kierunek zależności pomiędzy analizowanymi zmiennymi. Zmienne mogą być powiązane pomiędzy sobą za pomocą zależności funkcyjnej. W przypadku dwuwymiarowym każdej wielkości zmiennej niezależnej odpowiada określona jednoznacznie wartość zmiennej zależnej. Analiza regresji umożliwia między innymi określenie siły i rodzaju wpływu jednej zmiennej na drugą lub określenie wartości jednej zmiennej wykorzystując założone lub znane wartości drugiej zmiennej [Stanisz 2007].

Oszacowana funkcja regresji liniowej w przypadku dwuwymiarowym jest postaci:

$$\hat{y}_i = ax_i + b, \tag{1}$$

gdzie *a* i *b* są odpowiednio estymatorami parametru strukturalnego przy zmiennej objaśniającej i wyrazu wolnego tej funkcji [Luszniewicz, Słaby 2008].

Różnicę pomiędzy rzeczywistymi wartościami zmiennej objaśnianej (y_i) oraz odpowiadającymi im teoretycznymi wartościami tej zmiennej ($\hat{y}_i$) nazywamy składnikami resztowymi (e_i) modelu funkcji regresji. Wyznaczone są one według wzoru [Kot i in. 2007]:

$$e_i = y_i - \hat{y}_i . \tag{2}$$

Zanurzanie funkcji regresji liniowej zostało wprowadzone przez Hubert i Rousseeuw [1999]. Metoda oparta na tej koncepcji pozwala na oszacowanie parametrów strukturalnych równania regresji liniowej, która jest najbardziej zanurzona w stosunku do danych empirycznych oraz na określeniu stopnia jej zanurzenia w tym zbiorze. Estymator najbardziej zanurzonej regresji, może być wykorzystywane w przypadku występowania obserwacji nietypowych. Posługując się odpowiednim algorytmem, oprócz estymacji parametrów strukturalnych regresji liniowej, możliwe jest wyznaczenie wartości zanurzania modelu regresji w analizowanym zbiorze danych [Hubert, Rousseeuw 1999]. Wykorzystanie zanurzania w regresji liniowej przedstawione zostało między innymi w pracach Kobylińskiej [2011], Van Aelst, Rouusseuw [2000], Van Aelst i in. [2002]. W pracy definicje podane zostały dla przypadku dwuwymiarowego.

Niech P_n^2 określa zbiór dwuwymiarowy o liczebności *n* oraz niech (x_i, y_i) dla $i = 1, 2, ..., n$ będzie obserwacją należącą do P_n^2. Funkcję regresji liniowej $\hat{y}_i = ax_i + b$ nazywamy niedopasowaną do zbioru P_n^2, jeżeli istnieje taka liczba rzeczywista $\upsilon_y = \upsilon$, która nie pokrywa się z żadną wartością x_i, czyli $x_i \neq \upsilon$, że spełnione są następujące warunki

$e_i < 0$ dla każdego $x_i < \upsilon$ i $e_i > 0$ dla każdego $x_i > \upsilon$

lub

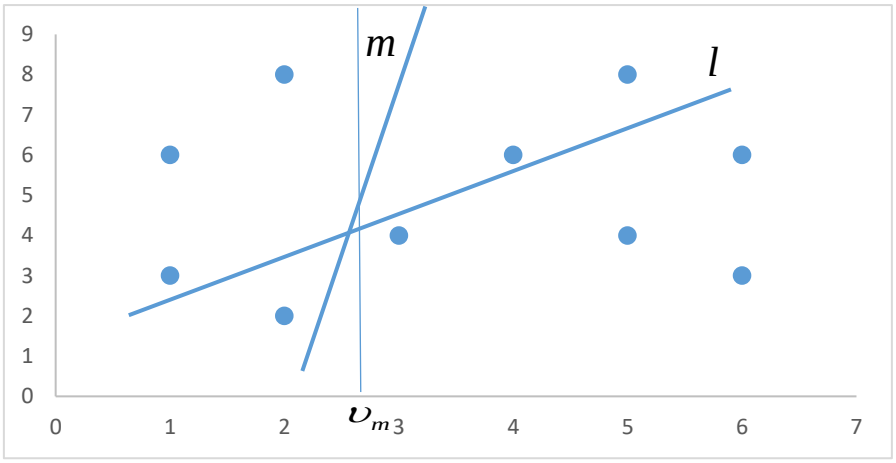
$e_i > 0$ dla każdego $x_i < \upsilon$ i $e_i < 0$ dla każdego $x_i > \upsilon$.

Zanurzaniem funkcji regresji liniowej $rzan(\hat{y}_i, P_n^2)$ w zbiorze dwuwymiarowym P_n^2 określamy najmniejszą liczbę obserwacji zbioru P_n^2, które powinny zostać wyeliminowane ze zbioru P_n^2, żeby funkcja $\hat{y}_i$ stała się niedopasowana do danego zbioru. Należy zauważyć, że jeżeli funkcja $\hat{y}_i$, położona jest powyżej lub poniżej wszystkich obserwacji danego zbioru, zawsze będzie niedopasowane do danych empirycznych [Hubert, Rousseeuw 1998].

Na rysunku 1 przedstawiono wykres korelacyjny zbioru P_{10}^2 o liczebności 10. Umieszczone zostały dwie funkcje regresji, spośród których *m* jest

niedopasowana do danych tego zbioru. Zauważyć można, że w przypadku tej funkcji istnieje liczba υ_m taka, że możemy obrócić funkcję *m* do pozycji pionowej w ten sposób, że nie przejdzie ona przez żaden punkt P_{10}^2. W przypadku prostej *l*, $rzan(m, P_{10}^2) = 4$, gdyż należy usunąć cztery punkty, żeby prosta *l* stała się niedopasowana do zbioru P_{10}^2. Inaczej mówiąc, zanurzanie funkcji regresji jest równe najmniejszej liczbie reszt, które muszą zmienić znak, żeby prosta regresji stała się niedopasowana do danych empirycznych.

Rysunek 1. Wykres korelacyjny zbioru P_{10}^2 oraz dwie funkcje, gdzie *m* jest niedopasowana do zbioru P_{10}^2



Źródło: opracowanie własne na podstawie Hubert i Rousseeuw [1999]

Maksymalne zanurzanie funkcji regresji w zbiorze P_n^2 spełnia nierówność $\max rzan(\hat{y}_i, P_n^2) \geq \lceil n/3 \rceil$,gdzie $\lceil A \rceil$ jest częścią całkowita liczby *A*. Dla każdego $P_n^2 \subset R^2$ zachodzi $\lceil n/3 \rceil \leq rzan(\hat{y}_i, P_n^2) \leq n$. Jeżeli wszystkie punkty P_n^2 leżą na prostej regresji $\hat{y}_i$, dla $i = 1, 2, ..., n$, wtedy $rzan(\hat{y}_i, P_n^2) = \max rzan(\hat{y}_i, P_n^2) = n.$ Własności dotyczące zanurzania regresji liniowej omówione zostały między innymi w pracy Hubert i Rousseeuw [1998].

PRZYKŁAD EMPIRYCZNY

W celu przedstawienia zastosowania metody zanurzania regresji liniowej wykorzystano dane liczbowe dwuwymiarowe. Do oszacowania funkcji regresji wykorzystano metodę najmniejszych kwadratów (obliczenia przeprowadzono

z wykorzystaniem pakietu STATISTICA PL) oraz metodę zanurzania regresji liniowej, dla której obliczenia przeprowadzono wykorzystując pakiet środowiska R „*DepthProc*”[1]. W skład pakietu wchodzi zestaw procedur statystycznych opartych na koncepcji zanurzania obserwacji w próbie. Informacje na temat pakietu oraz możliwości jego zastosowania w przypadku wielowymiarowym zamieszczone zostały między innymi w pracy Kosiorowskiego i Zawadzkiego [2014].

Na rysunkach 2-5 przedstawiono dane liczbowe dwuwymiarowe przyjmując za każdym razem inną konfigurację zbiorów (zbiory $P_{30}^2(xy)$, $P_{30}^2(x_1y_1)$, $P_{30}^2(x_4y_4)$, $P_{30}^2(x_5y_5)$). Dla każdego z tych zbiorów oszacowane zostały dwie funkcje regresji liniowej wykorzystując dane metody. Równania funkcji regresji zamieszczone zostały w tabeli 1.

Tabela 1. Oceny równań regresji z wykorzystaniem danych metod

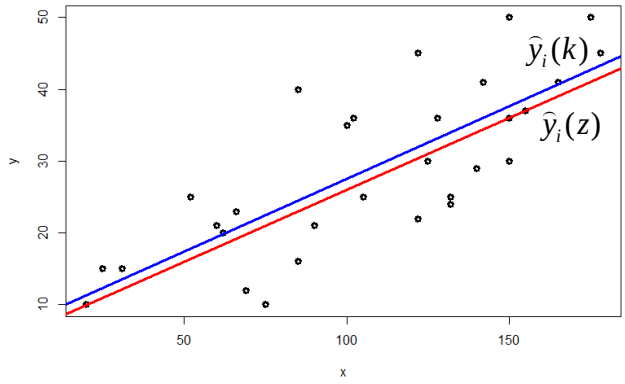
	Metoda zanurzania regresji	Metoda najmniejszych kwadratów
Ocena funkcji regresji	$\hat{y} = 6{,}00 + 0{,}20x$	$\hat{y} = 7{,}24 + 0{,}20x$
	$\hat{y}_1 = 6{,}00 + 0{,}20x_1$	$\hat{y}_1 = 11{,}59 + 0{,}15x_1$
	$\hat{y}_4 = 10{,}89 + 0{,}17x_4$	$\hat{y}_4 = 46{,}36 - 0{,}07x_4$
	$\hat{y}_5 = 8{,}67 + 0{,}20x_5$	$\hat{y}_5 = 7{,}98 + 0{,}32x_5$

Źródło: opracowanie własne

Nadmienić można, że zbiór danych $P_{30}^2(xy)$ (rysunek 2) tworzy zwartą figurę o kształcie pozwalającym aproksymować ją wykorzystując funkcję regresji liniowej. Zauważyć można, że funkcje regresji oszacowane dla tego zbioru położone są „blisko” siebie. Wynika to z faktu, że nie występują w tym przypadku obserwacje nietypowe, które mogą przyczynić się do zmiany położenia funkcji regresji. Współczynnik korelacji liniowej Pearsona wynosi dla zmiennych tego zbioru $r_{xy} = 0,78$ i jest to najwyższa wartość spośród współczynników korelacji wyznaczonych dla wszystkich rozpatrywanych zbiorów, które wynoszą odpowiednio $r_{x_1y_1} = 0,74$, $r_{x_4y_4} = 0,–0,18$ oraz $r_{x_5y_5} = 0,41$.

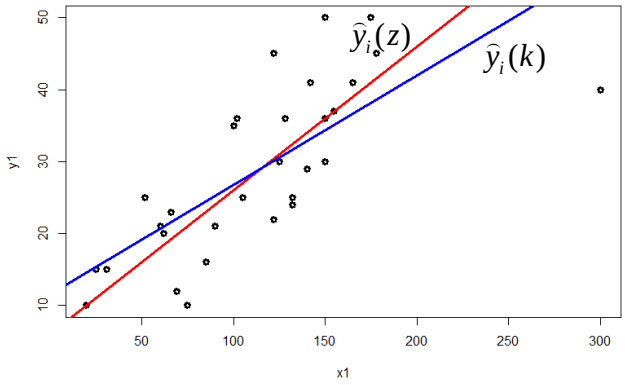
[1] https://cran.rstudio.com/web/packages/DepthProc/index.html

Rysunek 2. Wykres korelacyjny zbioru $P_{30}^{2}(xy)$ oraz funkcje regresji liniowej oszacowane metodą najmniejszych kwadratów ($\widehat{y}_i(k)$) oraz metodą maksymalnego zanurzania regresji ($\widehat{y}_i(z)$)



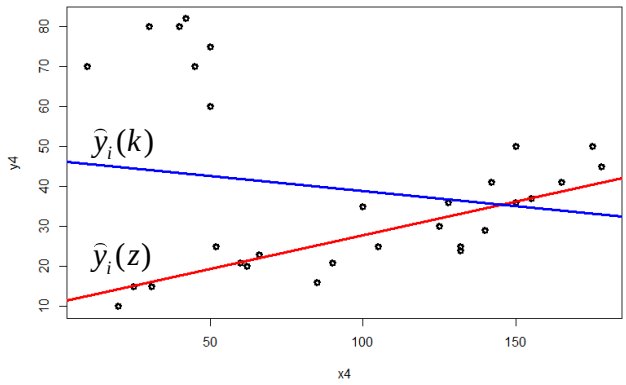
Źródło: opracowanie własne

Rysunek 3. Wykres korelacyjny zbioru $P_{30}^{2}(x_1y_1)$ oraz funkcje regresji liniowej oszacowane metodą najmniejszych kwadratów ($\widehat{y}_i(k)$) oraz metodą maksymalnego zanurzania regresji ($\widehat{y}_i(z)$)



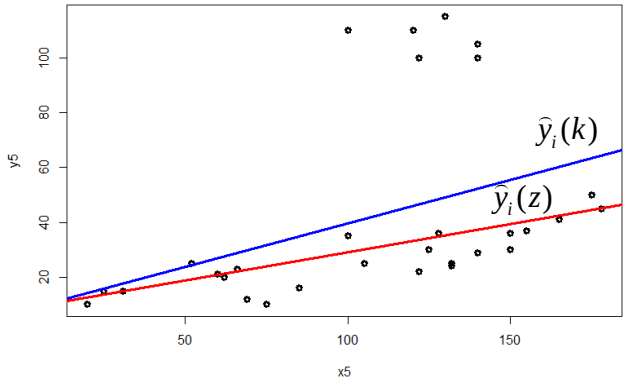
Źródło: opracowanie własne

Rysunek 4. Wykres korelacyjny zbioru $P_{30}^{2}(x_4 y_4)$ oraz funkcje regresji liniowej oszacowane metodą najmniejszych kwadratów ($\widehat{y}_i(k)$) oraz metodą maksymalnego zanurzania regresji ($\widehat{y}_i(z)$)



Źródło: opracowanie własne

Rysunek 5. Wykres korelacyjny zbioru $P_{30}^{2}(x_5 y_5)$ oraz funkcje regresji liniowej oszacowane metodą najmniejszych kwadratów ($\widehat{y}_i(k)$) oraz metodą maksymalnego zanurzania regresji ($\widehat{y}_i(z)$)



Źródło: opracowanie własne

Na rysunkach 3-5 zauważyć można występowanie obserwacji nietypowych, które nie pasują do konfiguracji rozpatrywanych zbiorów danych. Zmieniają one charakter zależności pomiędzy zmiennymi [Zeliaś 1996]. Występowanie w zbiorze

$P_{30}^2(x_1y_1)$ obserwacji, która jest odstająca w kierunku osi OX (rysunek 3), spowodowała zmianę nachylenia funkcji regresji, oszacowanej klasyczną metodą najmniejszych kwadratów. Regresja liniowa oszacowana z wykorzystanie metody opartej na koncepcji zanurzania obserwacji w próbie, nie zmieniła swojego położenia. Oceny parametrów strukturalnych funkcji oszacowanej tą metodą są takie same jak w przypadku zbioru $P_{30}^2(xy)$.

Na rysunku przedstawiającym zbiór $P_{30}^2(x_4y_4)$ (rysunek 4) zauważyć można występowanie podzbioru punktów nietypowych, którym odpowiadają niskie wartości zmiennej X oraz wysokie wartości drugiej zmiennej. Spowodowało to zmianę siły i kierunku zależności pomiędzy analizowanymi zmiennymi. Ocena współczynnika kierunkowego funkcji regresji liniowej oszacowanej metodą najmniejszych kwadratów jest ujemna (tabela 1). W przypadku wykorzystania metody najbardziej zanurzonej regresji nadal przyjmuje on wartość dodatnią. Zauważyć można, że występowanie podzbioru obserwacji nietypowych, nie ma wpływu na zmianę kierunku regresji liniowej w tym przypadku. W zbiorze $P_{30}^2(x_5y_5)$ (rysunek 5) występuje podzbiór obserwacji nietypowych ze względu na wysokie wartości zmiennej Y. Funkcja regresji oszacowana klasyczną metodą najmniejszych kwadratów umieszczona jest powyżej funkcji oszacowanej drugą metodą.

PODSUMOWANIE

Przedstawione rozważania prowadzą do wniosku, że dużym ograniczeniem, które spotykamy podczas szacowania parametrów funkcji regresji, jest problem występowania obserwacji nietypowych**.** W pracy przedstawiono użyteczność metod opartych na zanurzaniu obserwacji w próbie w szacowaniu parametrów strukturalnch funkcji regresji liniowej w przypadku występowania w zbiorze danych obserwacji znacznie odbiegających od pozostałych.

Wykorzystanie metody zanurzania regresji umożliwia oszacowanie regresji, dla większości punktów zbiorów danych. Na oceny parametrów regresji nie ma wpływu występowanie obserwacji nietypowych, modele regresji wykazują znacznie bliższe położenie w stosunku do większości punktów danych zbiorów. Estymator uzyskany z wykorzystaniem tej metody wykazuje się niska wrażliwością na zmiany wartości obserwacji w zbiorze danych [Hubert, Rousseeuw 1998].

BIBLIOGRAFIA

Andersen R. (2008) Modern Methods for Robust Regression. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-152.

Barnett V., Lewis T. (1978) Outliers in Statistical Data. Wiley and Sons, New York.

Jajuga K. (1993) Statystyczna analiza wielowymiarowa. PWN, Warszawa.

Hampel F. (2000) Robust Inference, Research Report 93. Seminar für Statistik, ETH Zürich. To appear: Encyclopedia of Environmetrics, Wiley. https://www.researchcollection.ethz.ch/bitstream/handle/20.500.11850/145174/eth-24068-01.pdf?sequence=1.

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

Kobylińska M. (2011) Zanurzanie w regresji liniowej. Metody Ilościowe w Badaniach Ekonomicznych, Warszawa, 12(2), 202-209.

Kosiorowski D. (2012) Statystyczne funkcje głębi w odpornej analizie ekonomicznej. Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Kosiorowski D., Zawadzki Z. (2014) DepthProc:An R Package for Robust Exploration of Multidimensional Economic Phenomene. http://arxiv.org/pdf/1408.4542.pdf .

Kot S., Jakubowski J., Sokołowski A. (2007) Statystyka. Podręcznik dla studentów ekonomicznych. Difin, Warszawa.

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

Stanisz A., (2007) Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Modele liniowe i nieliniowe, StatSoft, Kraków.

Rousseeuw P., Hubert M. (1999) Regression Depth. Journal of the American Statistical Association, 94, 388-402.

Rousseeuw P., Leroy A. (1987) Robust Regression and Outlier Detection. Wiley, New York.

Van Aelst S., Rouusseuw P. J. (2000) Robustness of Deepest Regression. Journal of Multivariate Analysis, 73, 82-106.

Van Aelst S., Rouusseuw P. J., Hubert M., Struyf A. (2002) The Deepest Regression method. Journal of Multivariate Analysis, 81, 138-166.

Zeliaś A. (1996) Metody wykrywania obserwacji nietypowych badaniach ekonomicznych. Wiadomości Statystyczne, 8, 16-27.

https://cran.rstudio.com/web/packages/DepthProc/index.html [dostęp 10.08.2018]

REGRESSION DEPTH METHOD FOR UNTYPICAL OBSERVATIONS

Abstract: This paper presents the application of the regression maximum depth for the estimation of linear regression function structural elements. For two-dimensional sets including untypical observations, regression functions were developed using the classical least squares method and a method based on the concept of observation depth measure in a sample. The effect of untypical observations on the estimated models has been noted.

Keywords: linear regression function, classical least squares method, observation depth measure in a sample, linear regression depth

JEL classification: C18, C40