UNIWERSYTET SZCZECIŃSKI
WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA

STUDIA I PRACE WYDZIAŁU NAUK EKONOMICZNYCH I ZARZĄDZANIA NR 36, T. 1

Barbara Batóg[*], **Jacek Batóg**[**]
Uniwersytet Szczeciński

ANALIZA WPŁYWU OBSERWACJI NIETYPOWYCH NA WYNIKI MODELOWANIA REGIONALNEJ WYDAJNOŚCI PRACY

STRESZCZENIE

W przeprowadzonym badaniu analizowano wpływ obserwacji nietypowych na modelowanie związków między regionalną wydajnością pracy a nakładami na innowacje w przemyśle i usługach oraz poziomem przedsiębiorczości w Polsce w latach 2002–2010. Uzyskane wyniki wskazują na stały wzrost poziomu przy jednoczesnym wzroście zróżnicowania wydajności pracy. W przypadku rozpatrywania wpływu nakładów na innowacje na wydajność pracy zidentyfikowano jedną obserwację nietypową (województwo mazowieckie) o charakterze dźwigni, natomiast w modelu z poziomem przedsiębiorczości wystąpiły dwie obserwacje wpływowe (województwa mazowieckie i zachodniopomorskie).

Słowa kluczowe: obserwacje nietypowe, regionalna wydajność pracy, nakłady na innowacje

Wstęp

O znaczącym wpływie większości obserwacji nietypowych (ang. *outliers*) na wyniki prowadzonych badań nie trzeba nikogo przekonywać. Dyskusyjna może

[*] Adres e-mail: barbara.batog@wneiz.pl

[**] Adres e-mail: batog@wneiz.pl

być tylko ocena rodzaju tych obserwacji oraz istotności ich wpływu na rezultaty modelowania ekonometrycznego. Rozważania na ten temat można znaleźć w wielu pracach [np. Barnett, Lewis 1994; Hawkins 1980]. Obszernej dyskusji doczekały się również rozważania dotyczące podstawowych przyczyn powstawania obserwacji nietypowych [Walfish 2006].

Ponieważ automatyczna eliminacja obserwacji uznanej za nietypową powoduje brak możliwości analizy przyczyny jej występowania, kluczową rolę odgrywa w tym przypadku umiejętność określenia charakteru obserwacji nietypowych, wśród których rozróżnia się najczęściej: *univariate outlier*, *regression outlier* (*vertical outlier*), *leverage* (dźwignia) oraz *influence* (wpływ) [Andersen 2008].

W tekście – na przykładzie modeli opisujących kształtowanie się regionalnych zmian wydajności pracy – zweryfikowana zostanie hipoteza badawcza mówiąca o istotnym wpływie obserwacji nietypowych na wyniki estymacji parametrów strukturalnych oraz jakość rozważanych modeli. Podstawowym celem analizy jest identyfikacja nietypowych obserwacji (regionów) w modelach wydajności pracy oraz ocena ich wpływu na proces modelowania tego zjawiska. Dodatkowym celem jest ocena wpływu nakładów na innowacje oraz poziomu przedsiębiorczości na wydajność pracy w ujęciu regionalnym.

Wydajność pracy uznawana jest za jeden z najważniejszych czynników decydujących w długim okresie o rozwoju, a w konsekwencji również o dobrobycie danego kraju lub regionu. Powszechnie znany jest ciąg przyczynowy: wzrost wydajności pracy, wzrost płac, rozwój społeczno-gospodarczy, poprawa jakości życia, wzrost dobrobytu. Zjawisko to odgrywa również znaczącą rolę w kształtowaniu przepływów siły roboczej oraz poziomu inwestycji. Różnice w poziomach wydajności pracy prowadzą do zróżnicowania dochodów *per capita*, przyczyniając się do dywergencji dochodowej, zwłaszcza w ujęciu regionalnym [zob. Batóg, Batóg 2008, 59–69].

Wybrane metody identyfikacji obserwacji nietypowych

Obszerną charakterystykę metod identyfikacji obserwacji nietypowych zawierają m.in. prace Belsleya, Kuha i Welscha, Ben-Gala, Ampanthonga oraz Williamsa i innych [Belsley, Kuh, Welsch 1980; Ben-Gal 2005, 3–12; Ampanthong 2009, Wil-

liams, Baxter, He, Hawkins, Gu 2002]. Do najpopularniejszych z nich i zastosowanych w niniejszym artykule należą:

– metoda oparta na wartościach h_i, określanych mianem *hat values*, które są miarą dźwigni (*leverage*) i pozwalają ocenić, w jakim stopniu wartość zmiennej niezależnej dla danej obserwacji odbiega od wartości średniej tej zmiennej:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

gdy $h_i > 2\bar{h}$ – obserwacja jest uznawana za nietypową;

– metoda standaryzowanych reszt:

$$e_i' = \frac{e_i}{S_e\sqrt{1-h_i}}$$

dla $|e_i'| > 2$ – obserwacja jest uznawana za nietypową;

– metoda studentyzowanych reszt:

$$e_i^* = \frac{e_i}{S_{e(-i)}\sqrt{1-h_i}} \sim t\,(n\text{-}k\text{-}2)$$

gdy $|e_i^*| > 2$ – obserwacja jest uznawana za nietypową;

– metoda DFBETAs (*difference of betas*), w której wartości D_{ij} stanowią miarę wpływu (*influence*) i pozwalają ocenić różnicę między wartościami ocen uzyskiwanymi dla regresji przy pełnym n oraz regresji z usuniętą wartością nietypową i:

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \text{ dla } i = 1, 2, \ldots, n \text{ oraz } j = 0, 1, \ldots, k$$

jeżeli $\frac{|D_{ij}|}{S_{-i}(\hat{\beta}_j)} \cdot \frac{2}{\sqrt{n}}$ – obserwacja uznawana jest za wpływową;

– metoda wykorzystująca odległość Cooka (D_i), która to miara w odróżnieniu od miary D_{ij} pozwala ocenić wpływ danej obserwacji na wszystkie oceny parametrów strukturalnych jednocześnie:

$$D_i = \frac{e_i'^2}{k+1} \cdot \frac{h_i}{1-h_i}$$

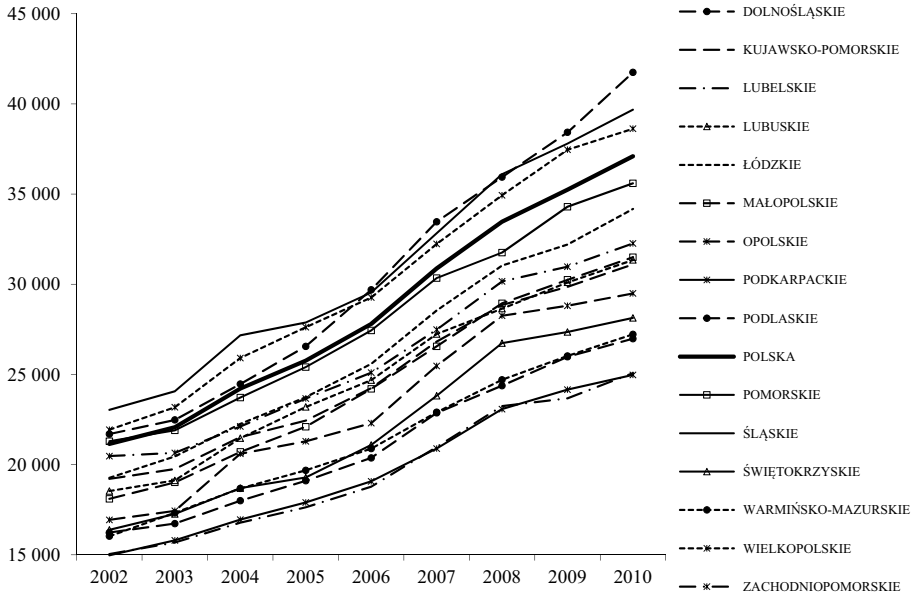
gdzie: pierwszy czynnik mierzy wpływ *vertical outlier*, a drugi efekt dźwigni, dla $D_i > \frac{4}{n-k-1}$ – obserwacja uznawana jest za wpływową;

– metody: DFFITS$_i$ (*difference of fits*), *partial regression plots*, *quantile comparison plots for studentized residuals*, Atkinson's Modified Cook's Statistics [Chatterjee, Hadi 1988; Rousseeuw, Leroy 1987].

Wyniki badań empirycznych

Wydajność pracy dla poszczególnych województw Polski została wyrażona przez zmienną PKB na jednego mieszkańca (zob. rys. 1). Obserwując jej kształtowanie się w latach 2002–2010 można zauważyć trzy prawidłowości. Pierwszą z nich jest znacząca przewaga województwa mazowieckiego w stosunku do wszystkich pozostałych województw (w 2002 r. PKB *per capita* kształtował się w tym województwie na poziomie 32 731 zł, a w 2010 r. przyjął wartość 60 359 zł, przewyższając drugie w kolejności województwo odpowiednio o 42,1% oraz 44,6%) oraz kształtowanie się wydajności pracy w województwach: dolnośląskim, wielkopolskim i śląskim powyżej średniej krajowej w całym badanym okresie. Drugą jest stały wzrost poziomu wydajności pracy, którego przeciętna wartość w ujęciu nominalnym była wyższa w 2010 r. w porównaniu do 2002 r. o 75,4%. Trzecią natomiast jest wzrost regionalnego zróżnicowania wydajności pracy mierzonego wartością współczynnika zmienności losowej (wzrost z poziomu 21% do 25%), co potwierdza często spotykany pogląd, że konwergencji dochodowej zachodzącej w skali krajów Unii Europejskiej towarzyszy dywergencja w ujęciu regionalnym [Batóg 2010].

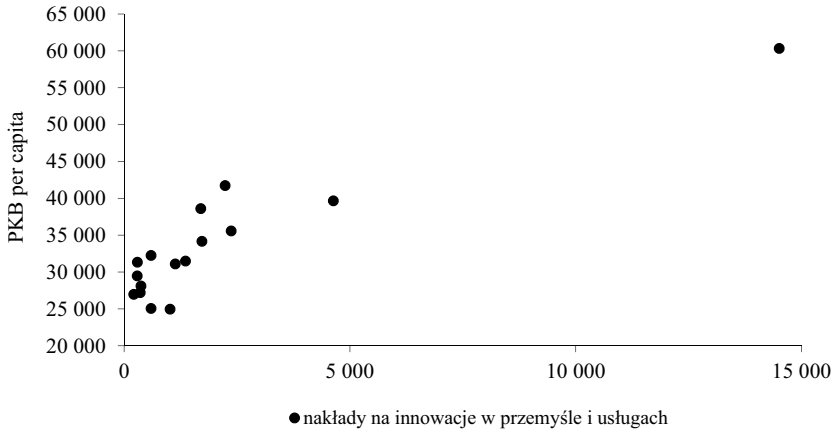
Rys. 1. Regionalny PKB na jednego mieszkańca w latach 2002–2010 – bez województwa mazowieckiego (zł)



Źródło: opracowanie własne na podstawie danych BDL GUS.

Na rysunku 2 przedstawiono w postaci graficznej zależności między wielkością PKB *per capita* oraz poziomem nakładów na innowacje w przemyśle i usługach według województw w 2010 r.

Rys. 2. PKB na jednego mieszkańca (zł) na tle nakładów na innowacje w przemyśle i usługach (mln zł) dla polskich województw w 2010 r



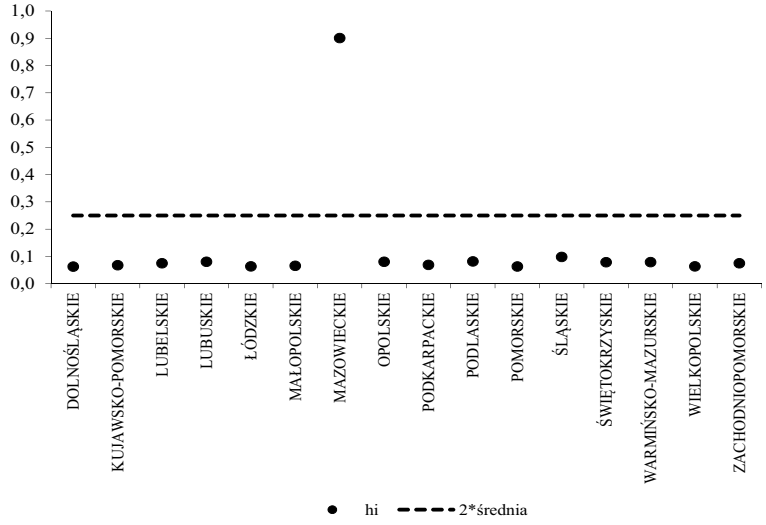
Źródło: opracowanie własne na podstawie danych BDL GUS.

Z powyższego diagramu korelacyjnego wynika, że istnieje silna korelacja dodatnia (współczynnik korelacji liniowej Pearsona $r_{xy} = 0{,}91$) oraz to, że jedna wartość znacząco różni się od pozostałych. Dotyczy ona województwa mazowieckiego. Wyniki estymacji parametrów strukturalnych modelu opisującego w ujęciu przekrojowym wpływ nakładów na innowacje X_i na wydajność pracy Y_i w 2010 r. (1) pozwalają stwierdzić, że wraz ze wzrostem zmiennej niezależnej o 1 mln zł wydajność pracy rosła przeciętnie o 2,28 zł (w modelu tym oraz w kolejnych, w nawiasach podane zostały średnie błędy szacunku parametrów strukturalnych):

$$\hat{y}_i = \underset{(1076,7)}{28894,5} + \underset{(0,27)}{2,28} X_i \,,\ R^2 = 0{,}835 \qquad (1)$$

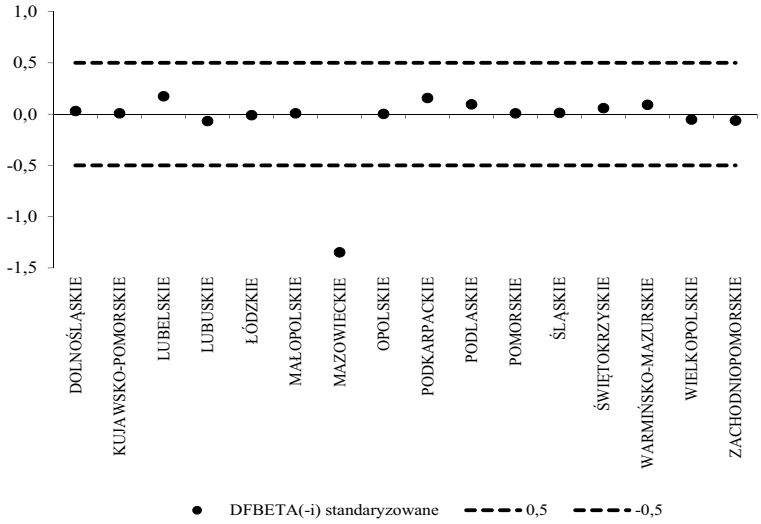
W celu rozstrzygnięcia, czy województwo mazowieckie powinno być uznane za obserwację nietypową, wykorzystano metody opisane w punkcie 1 (zob. rys. 3–5 oraz tab. 1). Ich wartości, poza podejściem opartym na resztach standaryzowanych i studentyzowanych wskazujących jako nietypową obserwację województwo dolnośląskie, potwierdzają wcześniejszy wniosek o nietypowości województwa mazowieckiego.

Rys. 3. Wartości h_i otrzymane dla modelu (1)



Źródło: opracowanie własne na podstawie danych BDL GUS.

Rys. 4. Wartości D_{ij} otrzymane dla modelu (1)



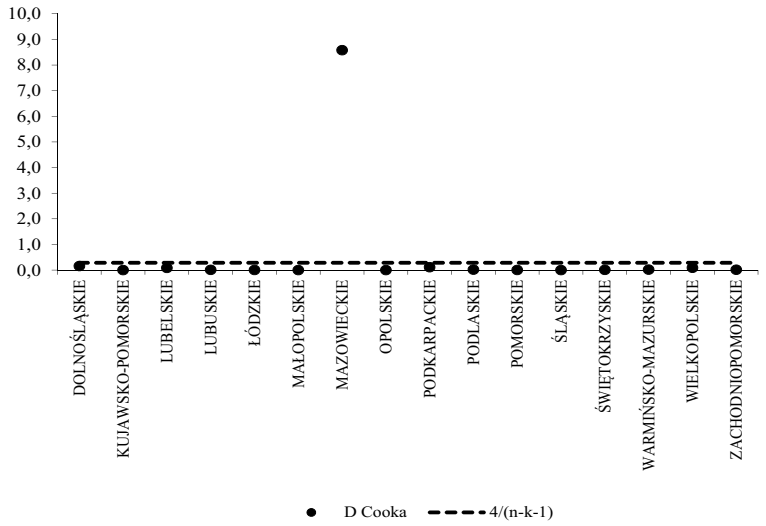
Źródło: opracowanie własne na podstawie danych BDL GUS.

Tabela 1. Reszty standaryzowane i studentyzowane w modelu (1)

Województwo	e_i	e_i'	e_i^*
łódzkie	1369,50	0,386	0,37
mazowieckie	–1592,09	–1,376	–1,34
małopolskie	–486,27	–0,137	–0,13
śląskie	228,00	0,065	0,06
lubelskie	–5167,04	–1,464	–1,53
podkarpackie	–6232,26	–1,760	–1,92
podlaskie	–2385,00	–0,678	–0,66
świętokrzyskie	–1602,73	–0,455	–0,44
lubuskie	1789,64	0,508	0,49
wielkopolskie	5875,68	1,655	1,78
zachodniopomorskie	2018,70	0,572	0,56
dolnośląskie	7762,03	2,185	2,59
opolskie	–53,06	–0,015	–0,01
kujawsko-pomorskie	–362,90	–0,102	–0,10
pomorskie	1307,10	0,368	0,36
warmińsko-mazurskie	–2469,29	–0,701	–0,69

Źródło: obliczenia własne na podstawie danych BDL GUS.

Rys. 5. Odległości D_i otrzymane dla modelu (1)



Źródło: opracowanie własne na podstawie danych BDL GUS.

Biorąc pod uwagę powyższe wyniki, oszacowano ponownie model (1), ale bez województwa mazowieckiego:

$$\hat{y}_i = \underset{(1360,0)}{27643,2} + \underset{(0,801)}{3,36}\, X_i,\ R^2 = 0,575 \tag{2}$$

Model (2), w porównaniu do jego wersji otrzymanej z wykorzystaniem wszystkich obserwacji, charakteryzuje się niższym dopasowaniem, a otrzymana ocena parametru stojącego przy nakładach na innowacje pozwala stwierdzić, że wraz ze wzrostem zmiennej niezależnej o 1 mln zł wydajność pracy rosła przeciętnie o 3,36 zł.

Oprócz nakładów na innowacje poziom wydajności pracy w poszczególnych województwach może być też uzależniony od poziomu przedsiębiorczości mierzonego liczbą podmiotów gospodarczych przypadajacych na tysiąc mieszkańców. Sugeruje to rysunek 6 przedstawiający zależność między tymi dwiema zmiennymi ($r_{xy} = 0,75$).

Rys. 6. PKB na jednego mieszkańca (zł) na tle liczby podmiotów gospodarczych na tysiąc mieszkańców dla polskich województw w 2010 r.

Źródło: opracowanie własne na podstawie danych BDL GUS.

Aby zweryfikować tę hipotezę, oszacowany został model (3) na podstawie danych przekrojowych z 2010 r.

$$\hat{y}_i = \underset{(9531,5)}{-6428,6} + \underset{(219,7)}{935,4}\, X_i,\ R^2 = 0,564 \tag{3}$$

gdzie:

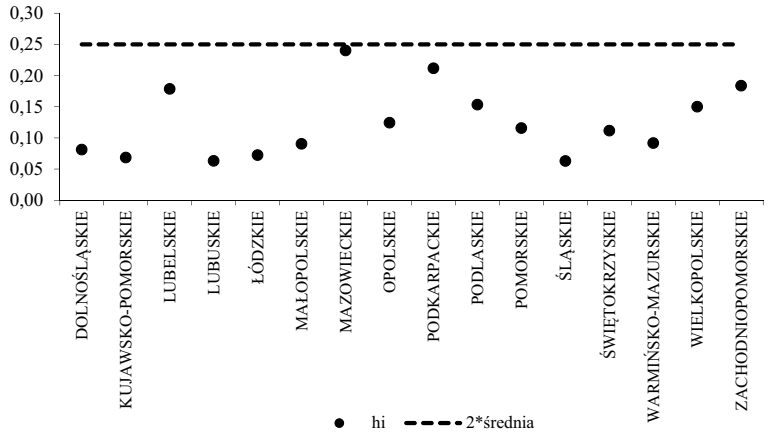
y_i – wydajność pracy,

X_i – poziom przedsiębiorczości mierzony liczbą podmiotów gospodarczych przypadających na tysiąc mieszkańców.

Uzyskane wyniki wskazują na niezbyt wysokie dopasowanie modelu do danych rzeczywistych, a ocena parametru przy zmiennej X_t informuje, że wydajność pracy wzrasta wraz ze wzrostem liczby firm przypadających na tysiąc mieszkańców o 1 o 935,4 zł.

Podczas przeprowadzonej identyfikacji obserwacji nietypowych (zob. rys. 7–9 oraz tab. 2) większość miar wskazała jako nietypowe województwo mazowieckie (poza wartością h_i) i zachodniopomorskie (poza wartością h_i i resztą standaryzowaną).

Rys. 7. Wartości h_i otrzymane dla modelu (3)



Źródło: opracowanie własne na podstawie danych BDL GUS.

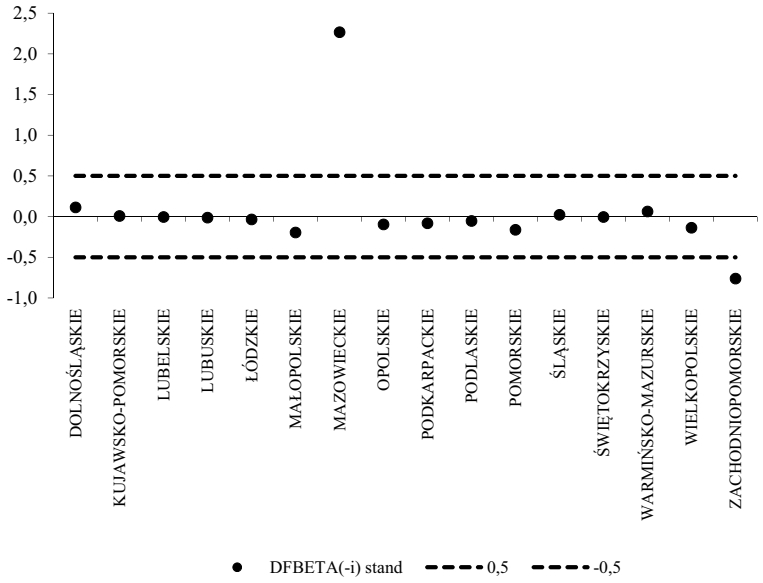
Tabela 2. Reszty standaryzowane i studentyzowane w modelu (3)

Województwo	e_i	e_i'	e_i^*
1	2	3	4
łódzkie	–1996,48	–0,347	–0,34
mazowieckie	16001,64	3,075	4,41
małopolskie	–6395,51	–1,123	–1,13

1	2	3	4
śląskie	5484,29	0,949	0,95
lubelskie	93,20	0,017	0,02
podkarpackie	1142,50	0,216	0,21
podlaskie	998,74	0,182	0,18
świętokrzyskie	122,25	0,022	0,02
lubuskie	–2919,70	–0,505	–0,49
wielkopolskie	–2535,05	–0,461	–0,45
zachodniopomorskie	–10225,83	–1,896	–2,07
dolnośląskie	4617,52	0,807	0,80
opolskie	2178,57	0,390	0,38
kujawsko-pomorskie	–580,24	–0,101	–0,10
pomorskie	–3910,06	–0,697	–0,68
warmińsko-mazurskie	–2075,85	–0,365	–0,35

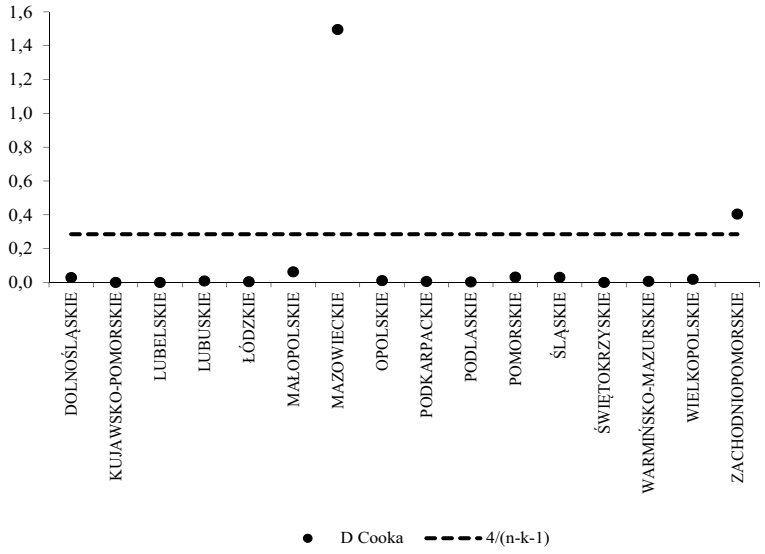
Źródło: obliczenia własne na podstawie danych BDL GUS.

Rys. 8. Wartości D_{ij} otrzymane dla modelu (3)



Źródło: opracowanie własne na podstawie danych BDL GUS.

Rys. 9. Odległości D_i otrzymane dla modelu (3)



Źródło: opracowanie własne na podstawie danych BDL GUS.

Wpływ wartości nietypowych na ocenę zależności między poziomem przedsiębiorczości i wydajnością pracy określić można przez porównanie modelu (3) z modelem (4), którego parametry został poddane estymacji na podstawie danych, z których usunięto obserwacje dla województwa mazowieckiego oraz zachodniopomorskiego:

$$\hat{y}_i = \underset{(5921,3)}{1253,,0} + \underset{(141,8)}{739,6} X_i,\ R^2 = 0,694 \tag{4}$$

Wraz z eliminacją wartości odstających widoczna jest znacząca poprawa jakości modelu oraz niewielki spadek wartości parametru strukturalnego przy zmiennej niezależnej.

Podsumowanie

W latach 2002–2010 obserwowany był w Polsce stały wzrost poziomu oraz zróżnicowania wydajności pracy w ujęciu regionalnym. Przeprowadzone badanie pozwoliło – z jednej strony – wykazać istotność wpływu nakładów na innowacje

w przemyśle i usługach oraz poziomu przedsiębiorczości na kształtowanie się regionalnej wydajności pracy, a z drugiej – zidentyfikować w procesie modelowania tego zjawiska obserwacje nietypowe o zróżnicowanym charakterze. W przypadku, gdy zmienną objaśniającą były nakłady na innowacje w przemyśle i usługach, za obserwację nietypową o charakterze dźwigni zostało uznane województwo mazowieckie. W modelu z poziomem przedsiębiorczości jako zmienną niezależną zastosowane metody wskazały jako obserwacje wpływowe dwa województwa: mazowieckie i zachodniopomorskie. Eliminacja wykrytych obserwacji nietypowych w obu powyższych przypadkach w znaczący sposób wpływała na poprawę uzyskanych wyników estymacji.

Literatura

Ampanthong P., Prachoom S. (2009), *A Comparative Study of Outlier Detection Procedures in Multiple Linear Regression*, w: *Proceedings of the International MultiConference of Engineers and Computer Scientists*, eds. S.I. Ao, O. Castillo, C. Douglas, D.D. Feng, J.-A. Lee, Hong Kong, Vol. I, IMECS 2009, March 18–20, s. 704–709.

Andersen R. (2008), *Modern Methods for Robust Regression*, Quantitative Applications in the Social Sciences 152, SAGE Publications, Los Angeles–London–New Delhi–Singapore.

Batóg J. (2010), *Konwergencja dochodowa w krajach Unii Europejskiej. Analiza ekonometryczna*, „Rozprawy i Studia" T. (DCCCLIV) 780, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Batóg J., Batóg B. (2008), *Analiza regionalnych zmian wydajności pracy w Polsce*, „Wiadomości Statystyczne", nr 6.

Barnett V., Lewis T. (1994), *Outliers in Statistical Data*, John Wiley & Sons, Chichester.

Belsley D.A., Kuh E., Welsch R.E. (1980), *Regression Diagnostics: Identyfing Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

Ben-Gal I. (2005), *Outlier detection*, w: *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, eds. O. Maimon, L. Rockach, Kluwer Academic Publishers, Boston.

Chatterjee S., Hadi A.S. (1988), *Sensitivity Analysis in Linear Regression*, John Wiley & Sons, New York.

Hawkins D. (1980), *Identification of Outliers*, Chapman and Hall, London.

Rousseeuw P.J., Leroy A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Walfish S. (2006), *A Review of Statistical Outlier Methods*, "Pharmaceutical Technology", November 2.

Williams G.J., Baxter R.A., He H.X., Hawkins S., Gu L. (2002), *A Comparative Study of RNN for Outlier Detection in Data Mining*, IEEE International Conference on Data-mining (ICDM'02), Maebashi City, CSIRO Technical Report CMIS-02/102.

ANALYSIS OF THE INFLUENCE OF OUTLIERS ON THE RESULS OF MODELLING OF REGIONAL LABOUR PRODUCTIVITY

Abstract

In the paper the Authors presented the analysis of the influence of outliers on results of econometric modelling of regional labour productivity. Innovation expenditures and a level of entrepreneurship were used as independent variables. Research was conducted for Polish voivodeships in 2002–2010. Two main types of outliers were distinguished: leverage and influence.

Translated by Barbara Batóg, Jacek Batóg

Keywords: outliers, regional labour productivity, innovation expenditures.
Kod JEL: C52, J24