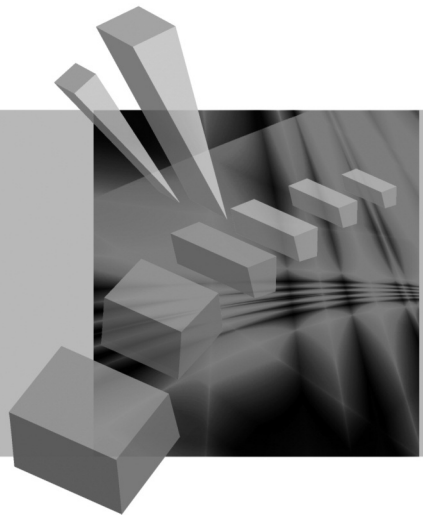
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Hanna Gruchociak
Uniwersytet Ekonomiczny w Poznaniu

KONSTRUKCJA ESTYMATORA REGRESYJNEGO DLA DANYCH O STRUKTURZE DWUPOZIOMOWEJ

Streszczenie: Głównym celem artykułu jest przedstawienie przydatności metodologii modelowania dwupoziomowego w szacowaniu wartości zmiennych społeczno-gospodarczych. W pierwszej części opracowania omówiona została idea konstrukcji estymatora dla danych o strukturze dwupoziomowej. W części drugiej przeprowadzone zostało badanie empiryczne, mające na celu zastosowanie opisanego estymatora do szacowania wskaźnika zatrudnienia w przekroju powiatów. Po dokonaniu oszacowań porównano wiarogodności estymatora uwzględniającego dwupoziomową strukturę danych oraz zwykłego estymatora regresyjnego. Przeprowadzona analiza wykazała istotną poprawę jakości oszacowań uzyskanych przy zastosowaniu modelowania dwupoziomowego.

Słowa kluczowe: modelowanie dwupoziomowe, struktura dwupoziomowa, część losowa, dojazdy do pracy, szacowanie zmiennych społeczno-gospodarczych.

1. Wstęp

Podstawowym celem artykułu jest ocena przydatności metodologii modelowania dwupoziomowego w szacowaniu charakterystyk społeczno-gospodarczych w przekroju terytorialnym. Zatem w pierwszej części opracowania omówiona zostanie idea konstrukcji estymatora dla danych o strukturze dwupoziomowej. W części drugiej przedstawione zostanie badanie empiryczne, celem którego będzie oszacowanie wskaźnika zatrudnienia w przekroju powiatów. Wskaźnik zatrudnienia jest jedną z charakterystyk rynku pracy, które z kolei zaliczają się do zmiennych opisujących sytuację społeczno-gospodarczą. Wysunięto hipotezę, że charakteryzuje się on strukturą dwupoziomową, której uwzględnienie poprawi precyzję estymacji.

Idea modelowania wielopoziomowego zrodziła się na początku lat 70., kiedy to badacze zwrócili uwagę na zróżnicowanie poziomu nauczania w klasach oraz szkołach. D. Lindley i A. Smith opracowali wtedy ogólne ramy dla badania zagnieżdżonych danych o złożonych strukturach błędów losowych (por. [Lindley, Smith 1972]).

Metodologia modelowania dwupoziomowego pozwala uwzględnić zależność jednostek badania należących do tej samej grupy. Ponadto, dzięki podzieleniu badanej populacji na grupy, możliwe jest wyjaśnienie części zmienności szacowanej ce-

chy za pomocą charakterystyk z drugiego poziomu. W przypadku estymowania zmiennych określonych na dwupoziomowej populacji zastosowanie omawianej metodologii w istotnym stopniu poprawia precyzję szacunku, pod warunkiem że zmienne te charakteryzują się dwupoziomową strukturą (por. [Goldstein 2003; Hox 2002; Raudenbush, Bryk 2002; Twisk Jos 2010]).

2. Założenia

Stosowanie metodologii modelowania dwupoziomowego jest uzasadnione tylko dla specyficznych zbiorowości oraz zmiennych. Badana populacja musi charakteryzować się dwupoziomową strukturą, co oznacza, że można ją podzielić na skończoną liczbę rozłącznych i pokrywających wszystkie jednostki pierwszego poziomu grup (inaczej jednostek drugiego poziomu). Również szacowana zmienna musi charakteryzować się dwupoziomową strukturą, co oznacza, że jej poziom powinien różnić się istotnie pomiędzy grupami. Zróżnicowanie to wynikać może z bezpośredniej zależności pomiędzy badaną zmienną a przynależnością jednostki pierwszego poziomu do grupy lub zależności badanej zmiennej oraz z podziału na grupy z pewną ukrytą, często niemierzalną zmienną. Kolejne założenie dotyczy normalności rozkładu szacowanej zmiennej, przy czym jego wartość oczekiwana różni się pomiędzy grupami, co jest konsekwencją dwupoziomowej struktury zmiennej, jednak wariancja powinna być stała w całej populacji.

3. Otrzymany estymator

W wyniku iteracyjnej konstrukcji modelu dwupoziomowego (por. [Goldstein 2003; Hox 2002; Raudenbush, Bryk 2002; Twisk Jos 2010]) otrzymano następującą funkcję regresji, pozwalającą na uwzględnienie dwupoziomowej struktury zmiennej objaśnianej:

$$Y_{ij} = \gamma_{00} + \sum_{q=1}^{Q}\left(\gamma_{0q}Z_{qj}\right) + e_{0j} + \\ + \sum_{p=1}^{P}\left(\left(\gamma_{p0} + \sum_{q=1}^{Q}\left(\gamma_{pq}Z_{qj}\right) + e_{pj}\right)X_{pij}\right) + r_{ij}\,, \qquad (1)$$

gdzie: n – liczebność całej próby,
J – liczba grup (liczba jednostek drugiego poziomu), $j = 1, \ldots, J$,
n_j – liczebność próby w j-tej grupie ($\sum_{j=1}^{J} n_j = n$),
Y_{ij} – wartość zmiennej objaśnianej dla *i-tej* obserwacji z j-tej grupy,
P – liczba zmiennych objaśniających z pierwszego poziomu,
X_{pij} – wartość p-tej zmiennej objaśniającej z pierwszego poziomu dla i-tej obserwacji z j-tej grupy, $p = 1, \ldots, P$,
Q – liczba zmiennych objaśniających z drugiego poziomu,
Z_{qj} – wartość q-tej zmiennej objaśniającej z drugiego poziomu dla j-tej jednostki drugiego poziomu, $q = 1,\ldots, Q$,

r_{ij} , e_{pj} – niezależne reszty dla jednostek pierwszego i drugiego poziomu,
γ_{pq} – parametry regresji, $p = 0,1,\ldots, P$, $q = 0, 1, \ldots, Q$.

Zatem estymator regresyjny dla danych o strukturze dwupoziomowej można zapisać wzorem:

$$\hat{Y}_{ij}^{D} = \gamma_{00} + \sum_{q=1}^{Q}(\gamma_{0q}Z_{qj}) + \sum_{p=1}^{P}\left((\gamma_{p0} + \sum_{q=1}^{Q}(\gamma_{pq}Z_{qj}))X_{pij}\right) \tag{2}$$

z błędem losowym:

$$e_{0j} + \sum_{p=1}^{P}(e_{pj} * X_{pij}) + r_{ij}. \tag{3}$$

Powyższe funkcje przedstawić można równoważnie w zapisie macierzowym:

$$\boldsymbol{Y^{D}} = diag[\boldsymbol{X}(\boldsymbol{\gamma Z} + \boldsymbol{E})\boldsymbol{A}] + \boldsymbol{R} \tag{4}$$

$$\boldsymbol{\hat{Y}^{D}} = diag(\boldsymbol{X\gamma ZA}) \tag{5}$$

z błędem losowym:

$$diag(\boldsymbol{XEA}) + \boldsymbol{R}\ , \tag{6}$$

gdzie:

$$\boldsymbol{Y}^{\boldsymbol{D}}_{n\times 1} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ \hline \vdots \\ Y_{1J} \\ \vdots \\ Y_{n_J J} \end{bmatrix}, \boldsymbol{X}_{n\times(P+1)} = \left[\begin{array}{cccc} 1 & X_{111} & \cdots & X_{P11} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n_1 1} & \cdots & X_{Pn_1 1} \\ \hline \vdots & \vdots & & \vdots \\ \hline 1 & X_{11J} & \cdots & X_{P1J} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n_J J} & \cdots & X_{Pn_J J} \end{array}\right],$$

$$\boldsymbol{\gamma}_{(P+1)\times(Q+1)} = \begin{bmatrix} \gamma_{00} & \gamma_{01} & \cdots & \gamma_{0Q} \\ \gamma_{10} & \gamma_{11} & \cdots & \gamma_{1Q} \\ \vdots & \vdots & & \vdots \\ \gamma_{P0} & \gamma_{P1} & \cdots & \gamma_{PQ} \end{bmatrix}, \boldsymbol{R}_{n\times 1} = \begin{bmatrix} r_{11} \\ \vdots \\ r_{n_1 1} \\ \hline \vdots \\ r_{1J} \\ \vdots \\ r_{n_J J} \end{bmatrix}$$

$$\boldsymbol{Z}_{(Q+1)\times J} = \begin{bmatrix} 1 & \cdots & 1 \\ Z_{11} & \cdots & Z_{1J} \\ \vdots & & \vdots \\ Z_{Q1} & \ldots & Z_{QJ} \end{bmatrix}, \boldsymbol{A}_{J\times n} = \begin{bmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \quad \ddots \quad \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ 1 & \cdots & 1 \end{bmatrix},$$

$$\boldsymbol{E}_{(P+1)\times J} = \begin{bmatrix} e_{01} & \cdots & e_{0J} \\ e_{11} & \cdots & e_{1J} \\ \vdots & & \vdots \\ e_{P1} & \ldots & e_{PJ} \end{bmatrix}$$

$$\boldsymbol{R}_{n\times 1}\sim N_n(\boldsymbol{0}_n;\sigma^2\boldsymbol{I}_n),\ \boldsymbol{E}_{(P+1)\times J}\sim N_{J\times(P+1)}\left(\boldsymbol{0}_{(P+1)\times J};\boldsymbol{I}_J\otimes\boldsymbol{T}\right),$$

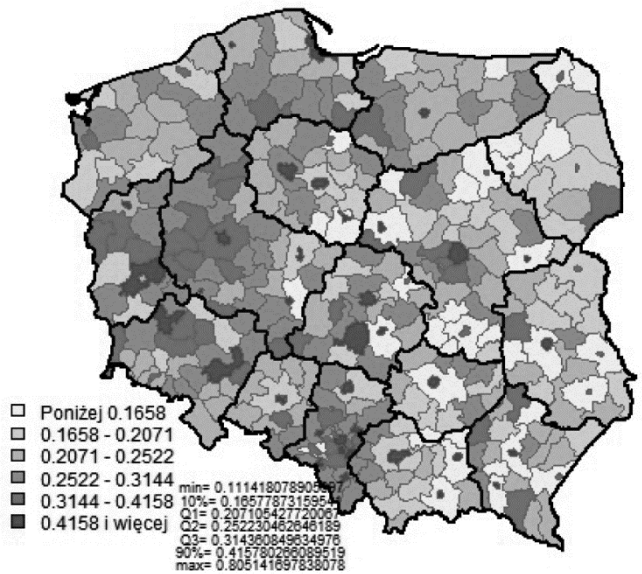
gdzie:

$$\boldsymbol{T}_{(P+1)\times(P+1)}=\begin{bmatrix}\tau_{00} & \cdots & \tau_{0P}\\ \vdots & \ddots & \vdots\\ \tau_{0P} & \cdots & \tau_{PP}\end{bmatrix}. \tag{7}$$

Rozwiązanie powyższego problemu sprowadza się do oszacowania współczynników funkcji regresji opisanej wzorami (2) i (4), zawartych w macierzy $\boldsymbol{\gamma}$. Współczynniki te szacowane są zgodnie z metodą największej wiarogodności.

4. Badanie empiryczne

Aktywność ekonomiczna ludności w przekroju terytorialnym zdeterminowana jest m.in. poziomem rozwoju gospodarczego regionu, w wyniku czego poziom względnych charakterystyk rynku pracy różni się pomiędzy województwami. Wysunięto przypuszczenie, że poziom wskaźnika zatrudnienia jako jednej z charakterystyk rynku pracy różni się istotnie pomiędzy województwami (por. rys. 1). Jeżeli hipoteza ta jest prawdziwa, zastosowanie metodologii modelowania dwupoziomowego mogłoby w istotny sposób poprawić precyzję oszacowań tej charakterystyki rynku pracy.



Rys. 1. Rozkład przestrzenny wskaźnika zatrudnienia, przekrój powiatów, Polska 2006

Źródło: opracowanie własne.

Za jednostki pierwszego poziomu przyjęto powiaty, zaś jako jednostki drugiego poziomu – województwa. Czas badania określono na rok 2006, ze względu na dostępność danych na temat dojazdów do pracy, wykorzystanych jako jedna ze zmiennych objaśniających.

W celu porównania oszacowano również parametry klasycznej regresji liniowej.

5. Zmienne zastosowane do analizy

Szacowaną zmienną jest wskaźnik zatrudnienia, zdefiniowany jako stosunek liczby osób zatrudnionych w ich głównym miejscu pracy (bez uwzględnienia podmiotów gospodarczych o liczbie pracujących do 9 osób) na terenie powiatu do liczby osób w wieku produkcyjnym, tj. od 15 roku życia do wieku emerytalnego, zamieszkujących na terenie tego powiatu. Wysoki poziom wskaźnika zatrudnienia związany jest z dużym nasileniem przyjazdów do pracy do danego powiatu oraz niskim poziomem wyjazdów do pracy z tego powiatu. Duży udział osób zatrudnionych w danym powiecie świadczy o jego dobrej sytuacji, co motywuje do podejmowania pracy na jego terenie. Z drugiej strony silne natężenie przyjazdów do pracy do danego powiatu, przy jednoczesnym niskim poziomie wyjazdów do pracy podejmowanych przez jego mieszkańców, powoduje zwiększanie się liczby osób zatrudnionych na jego terenie przy niezmienionej liczbie osób w wieku produkcyjnym, co jest równoznaczne ze wzrostem wskaźnika zatrudnienia. Jako pierwszą zmienną objaśniającą na poziomie powiatu przyjęto zatem intensywność dojazdów do pracy określoną jako stosunek salda dojazdów do pracy[1] do ludności w wieku produkcyjnym (w dalszych analizach X_1).

Ze względu na ścisły związek pomiędzy wskaźnikiem zatrudnienia i poziomem bezrobocia jako drugą zmienną objaśniającą na poziomie powiatu przyjęto natężenie bezrobocia zdefiniowane jako stosunek liczby bezrobotnych do ludności w wieku produkcyjnym (w dalszych analizach X_2).

Zgodnie z teorią Thunena wielkie miasta stymulują rozwój terenów je otaczających. W związku z tym w miarę zwiększania się odległości od ośrodków centralnych pogarsza się sytuacja gospodarcza powiatu, m.in. maleje wskaźnik zatrudnienia. Dlatego jako trzecią zmienną objaśniającą na poziomie powiatów przyjęto odległość od najbliższego wielkiego miasta (w dalszych analizach X_3). Zbiór wielkich miast zdefiniowano jako 24 największe ze względu na liczbę zatrudnionych miasta na prawach powiatu. Odległość w kilometrach od centroidu każdego powiatu od najbliższego ośrodka centralnego obliczono za pomocą pakietu *nlme* z programu R (por. [Bliese 2009]).

Jako pierwszą zmienną z poziomu województw wybrano wskaźnik zatrudnienia zdefiniowany analogicznie jak zmienna objaśniana, ale na zbiorowości województw (w dalszych analizach Z_1).

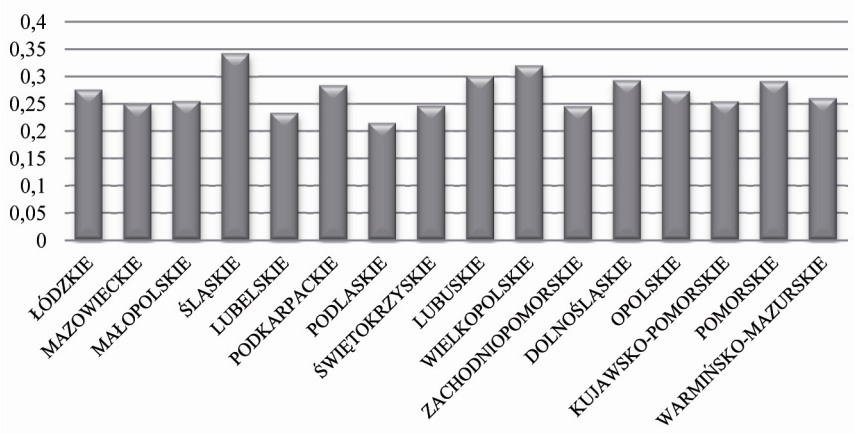
[1] Różnica pomiędzy liczbą osób wyjeżdżających do pracy a liczbą osób przyjeżdżających do pracy.

Ponieważ wyższe wykształcenie ułatwia znalezienie pracy, wysoki poziom wykształcenia ludności pociąga za sobą wysoką wartość wskaźnika zatrudnienia. Zatem jako drugą zmienną objaśniającą z poziomu województw przyjęto poziom wykształcenia, zdefiniowany jako udział osób z wyższym wykształceniem w ludności w wieku produkcyjnym (w dalszych analizach Z_2).

6. Weryfikacja dwupoziomowej struktury zmiennej objaśnianej

Jak stwierdzono powyżej, występują merytoryczne przesłanki zróżnicowania wskaźnika zatrudnienia w powiatach należących do różnych województw. Aby upewnić się, że zróżnicowanie to jest statystycznie istotne, przeprowadzono test analizy wariancji. Weryfikacji poddana została hipoteza zerowa o braku istotnych różnic w średnim poziomie zatrudnienia w powiatach pomiędzy województwami.

$$\begin{cases} \boldsymbol{H_0}: \boldsymbol{\mu_1 = \cdots = \mu_J} \\ \boldsymbol{H_1}: \sim \boldsymbol{H_0} \end{cases}.$$



Rys. 2. Przeciętny wskaźnik zatrudnienia w przekroju powiatów według województw, Polska 2006

Źródło: opracowanie własne na podstawie BDL.

Wyliczona została wartość statystyki *F*-Snedecora: $F = \frac{\frac{\Sigma_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2}{J-1}}{\frac{\Sigma_{j=1}^{J} \Sigma_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2}{n-J}} = 2{,}93$ oraz kwantyl rozkładu F-Snedecora dla obranego poziomu istotności 5% i stopni swobody równych odpowiednio liczbie województw pomniejszonej o jeden i liczbie powiatów pomniejszonej o liczbę województw: $F(0{,}95; 15; 350) = 0{,}48$. Ponieważ wartość wyliczonej statystyki jest znacznie większa od wartości krytycznej: $B = \{\underline{Y}: 2{,}9332 \geq 0{,}4802\}$, hipoteza zerowa zostaje odrzucona na korzyść hipotezy alternatywnej.

7. Estymator wskaźnika zatrudnienia

Następnie wyznaczono, zgodnie ze wzorem 2, estymator regresyjny dla danych o strukturze dwupoziomowej:

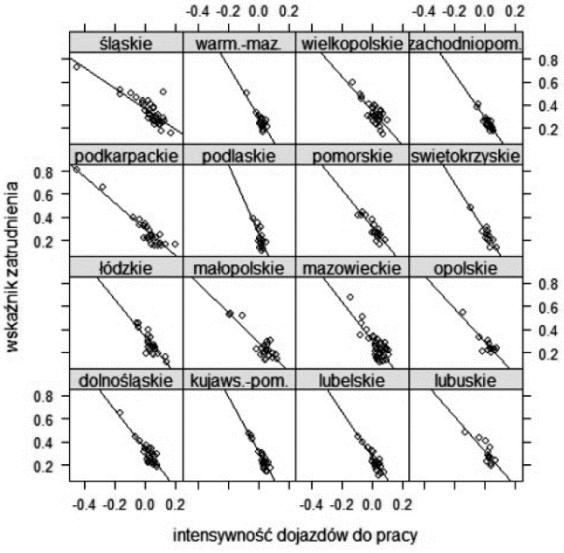
$$\hat{Y}_{ij}^{D} = 0{,}174 + 0{,}487 * Z_{1j} + 3{,}329 * Z_{2j} +$$
$$+\left(-1{,}380 + 4{,}935 * Z_{1j} - 105{,}816 * Z_{2j}\right) * X_{1ij} + (0{,}005 - 0{,}672 * Z_{2j}) *$$
$$* \ X_{2ij} - 0{,}0002 * X_{3ij}.$$

Analizując wyznaczone współczynniki, zauważyć można, że do oszacowania wskaźnika zatrudnienia na poziomie powiatu wliczana jest ok. połowa (0,487) wartości tego wskaźnika na poziomie województwa. Z kolei wzrost udziału osób z wykształceniem wyższym w ludności w wieku produkcyjnym na terenie województwa o 1% powoduje wzrost oszacowania wskaźnika zatrudnienia w powiatach należących do tego województwa o 0,0329. Współczynnik kierunkowy przy intensywności dojazdów do pracy jest ujemny, ponieważ duża liczba osób wyjeżdżających do pracy przy małej liczbie przyjazdów związana jest z niskim poziomem zatrudnienia w powiecie. Z przeprowadzonej analizy wynika, że nasilenie omawianej zależności różni się istotnie pomiędzy województwami (por. rys. 3a), dlatego też współczynnik przy omawianej zmiennej przyjmuje różne wartości dla różnych województw. Mianowicie dla województw o wysokim poziomie zatrudnienia (Z_{1j}) zależność pomiędzy wskaźnikiem zatrudnienia i intensywnością dojazdów do pracy w powiatach traci na sile, gdyż ujemny współczynnik kierunkowy powiększony zostaje o $4{,}935 * Z_{1j}$. Z kolei w województwach charakteryzujących się wysokim poziomem wykształcenia zależność pomiędzy zmienną objaśnianą a intensywnością dojazdów do pracy zyskuje na sile. Wynika to z większej gotowości do podejmowania wysiłku dojazdów do pracy wśród osób wykształconych, które nie zawsze są w stanie znaleźć odpowiadającą swoim kwalifikacjom pracę w miejscu zamieszkania. Ujemny współczynnik kierunkowy przy zmiennej X_{1ij} został pomniejszony o $105{,}816 * Z_{2j}$. Również w przypadku zależności pomiędzy wskaźnikiem zatrudnienia oraz poziomem bezrobocia w przekroju powiatów stwierdzono istotne zróżnicowanie jej charakteru pomiędzy województwami (por. rys. 3b). We wszystkich powiatach była to zależność odwrotnie proporcjonalna, jednak różniła się intensywnością. Zauważono, że w województwach o wysokim poziomie wykształcenia zależność pomiędzy wskaźnikiem zatrudnienia i poziomem bezrobocia zyskuje na sile. Analizując charakter zależności pomiędzy wskaźnikiem zatrudnienia oraz odległością od najbliższego wielkiego miasta, nie zauważono istotnych różnic pomiędzy województwami. Niezależnie od województwa powiaty położone bliżej ośrodków centralnych charakteryzowały się wyższym poziomem zatrudnienia. Uzasadnienie takiej zależności odwrotnie proporcjonalnej znaleźć można na przykład w teorii Thunena.

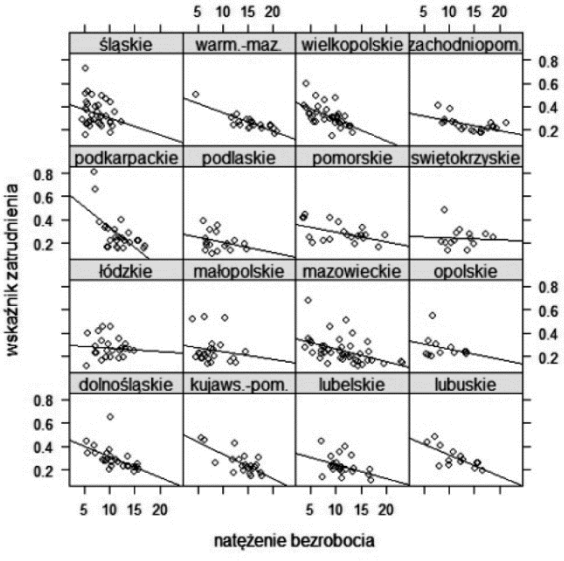
W celu porównań oszacowano również klasyczny estymator regresyjny:

$$\hat{Y}_{ij}^{K} = \ 0{,}389 - 1{,}199 * X_{1ij} - 0{,}006 * X_{2ij} - 0{,}0003 * X_{3ij}.$$

a. Linie regresji dla X_1



b. Linie regresji dla X_2



Rys. 3. Zależność pomiędzy zmienną objaśnianą i wybranymi zmiennymi objaśniającymi w przekroju powiatów, dla każdego z województw rozważana indywidualnie

Źródło: opracowanie własne.

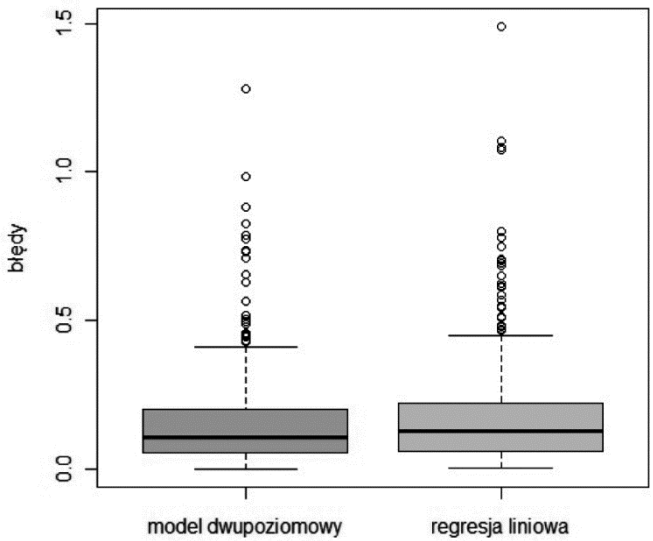
8. Porównanie obu modeli

Po dokonaniu oszacowań wskaźnika zatrudnienia za pomocą estymatora regresyjnego dla danych o strukturze dwupoziomowej oraz klasycznego estymatora regresyjnego porównano otrzymane wyniki. W tym celu obliczono logarytmy wiarogodności obu modeli i zweryfikowano za pomocą testu ilorazu wiarogodności istotną poprawę modelu uwzględniającego dwupoziomową strukturę danych w stosunku do modelu klasycznej regresji liniowej (por. [Harville 1974]). Na tej podstawie stwierdzono przewagę jakości dopasowania modelu dwupoziomowego istotną dla praktycznie dowolnego poziomu ufności (por. tab. 1). Wyznaczono również wartości kryteriów informacyjnych Akaike'a (por. [Sakamoto, Ishiguro, Kitagawa 1986]) oraz Bayesowskiego (por. [Schwarz 1978]). Otrzymane wyniki również wskazywały jednoznacznie na przewagę modelu dwupoziomowego (por. tab. 1).

Tabela 1. Porównanie jakości dopasowania do danych modelu klasycznej regresji liniowej oraz dwupoziomowego na podstawie wybranych kryteriów

Kryteria oceny jakości dopasowania:	*AIC*	*BIC*	*lnL*	$\hat{Y}_{ij}^{D}$ *vs* $\hat{Y}_{ij}^{K}$	
Klasyczna regresja liniowa	–1054,14	–1034,54	532,069	$\chi 2$	*p-value*
Model dwupoziomowy	–1161,46	–1110,69	593,728	123,319	<0,0001

Źródło: opracowanie własne.



Rys. 4. Błędy względne oszacowań wskaźnika zatrudnienia, przekrój powiatów, Polska 2006

Źródło: opracowanie własne.

Dla większości powiatów oszacowanie otrzymane przy zastosowaniu estymatora uwzględniającego dwupoziomową strukturę danych okazały się bliższe wartościom empirycznym niż oszacowania uzyskane w modelu klasycznej regresji liniowej. Zaowocowało to mniejszymi wartościami błędów względnych oszacowań, szczególnie niższą wartością ich mediany oraz kwartyla trzeciego. Również maksymalny błąd względny był niższy w przypadku uwzględnienia dwupoziomowej struktury szacowanej zmiennej (por. rys. 4).

9. Wnioski z przeprowadzonej analizy

Przedstawione w pracy badanie może być traktowane jako empiryczna weryfikacja przydatności metodologii modelowania dwupoziomowego w szacowaniu charakterystyk społeczno-gospodarczych w przekroju terytorialnym. Na podstawie przeprowadzonej analizy stwierdzono, że wskaźniki zatrudnienia w powiatach charakteryzują się dwupoziomową strukturą, wynikającą m.in. ze zróżnicowania poziomu rozwoju gospodarczego województw. Dzięki temu zastosowanie estymatora regresyjnego dla danych o strukturze dwupoziomowej pozwoliło uzyskać statystycznie istotną poprawę jakości szacunku, wyrażoną wiarogodnością modelu oraz poprawą precyzji wyznaczonych oszacowań.

Literatura

Bliese P., *Multilevel Modeling in R (2.3) A Brief Introduction to R, the Multilevel Package and the NLME Package*, Paul Bliese, 2009.

Goldstein H., *Multilevel Statistical Models*, 3rd edition, Edward Arnold, London 2003.

Harville D.A., *Bayesian inference for variance components using only error contrasts*, „*Biometrika*" 1974, no 61.

Hox J., *Multilevel Analysis. Techniques and Applications*, Lawrence Erlbaum Associates, Publishers, London 2002.

Lindley D., Smith A., *Bayes estimates for the linear model*, „Journal of the Royal Statistical Society" 1972, Series B, no 34.

Raudenbush S.W., Bryk A.S., *Hierarchical Linear Models. Applications and Data Analysis Methods*, Second Edition, Sage Publications, London_Thousand Oaks_New Delhi 2002.

Sakamoto Y., Ishiguro M., Kitagawa G., *Akaike Information Criterion Statistics*, D. Reidel Publishing Company, 1986.

Schwarz G., *Estimating the dimension of a model*, „Annals of Statistics" 1978, no 6.

Twisk Jos W.R., *Analiza wielopoziomowa – przykłady zastosowań. Praktyczny podręcznik biostatyki i epidemiologii*, Oficyna Wydawnicza SGH, Warszawa 2010.

CONSTRUCTION OF REGRESSION ESTIMATOR FOR TWO-LEVEL DATA

Summary: The main goal of this article is to demonstrate the usefulness of two-level modeling methodology for estimating the value of socio-economic variables. The first part treats of the idea of estimator construction for the two-level structure data. The second part describes the empirical studies which were conducted to show the application of described estimator to estimate the employment in the cross-section of counties. After the estimation the likelihood of the estimator which took into account the two-level structure was compared to a classical regression estimator. The analysis showed a significant improvement of the quality of estimates obtained using two-level modeling.

Keywords: two-level modeling, two-level structure, random effects, commuter routes, estimating the socio-economic variables.