PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 4 – 2012

HANNA GRUCHOCIAK

MOŻLIWOŚCI ZASTOSOWANIA MODELOWANIA DWUPOZIOMOWEGO W BADANIACH EKONOMICZNYCH

1. WPROWADZENIE

Głównym celem artykułu jest przedstawienie przydatności metodologii modelowania dwupoziomowego w zastosowaniach społeczno-ekonomicznych. Tak więc w pierwszej części opracowania przedstawione zostaną etapy komplikowania klasycznego modelu regresji liniowej oraz wybrane kryteria oceny poprawy jego dopasowania. W części drugiej przedstawiony zostanie przykład zastosowania opisanej metodologii do szacowania liczby osób pracujących w przekroju powiatów.

W pierwszej kolejności omówione zostaną etapy tworzenia funkcji regresji opartej na danych o strukturze dwupoziomowej. Punktem wyjścia będzie klasyczny model regresji liniowej nieuwzględniający dwupoziomowej struktury danych. W etapie końcowym zaprezentowany zostanie model uwzględniający wpływ losowego wyrazu wolnego oraz losowych współczynników, zależnych od zmiennych objaśniających z drugiego poziomu. Warto w tym miejscu zaznaczyć, że forma końcowego modelu dwupoziomowego jest zazwyczaj taka sama dla dostępnych w literaturze opracowań (por. Raudenbush, Bryk, 2002; Węziak, 2007; Bliese, 2012), jednak zaobserwować można różne podejścia w stopniowym komplikowaniu modelu. Według jednego z podejść zmienne objaśniające z drugiego poziomu zostają dołączone do modelu przed zmiennymi objaśniającymi z poziomu pierwszego (por. Klimanek, 2003; Raudenbush, Bryk, 2002). W niniejszym opracowaniu omówiono natomiast podejście, w którym najpierw dodaje się zmienne objaśniające z poziomu pierwszego (por. Bliese, 2012; Twisk, 2010). Taki sposób postępowania uznano za bardziej intuicyjny i lepiej odpowiadający sytuacjom rzeczywistym.

Stopniowe komplikowanie modelu ma na celu uniknięcie uwzględnianie losowego charakteru parametrów modelu, jeżeli nie poprawia w istotny sposób dopasowania modelu do konkretnych danych rzeczywistych. W związku z tym omówiono także kilka wybranych kryteriów pozwalających ocenić poprawę jakości oszacowania modelu w porównaniu z modelem prostszym (z poprzedniego etapu). W zastosowaniach praktycznych należy po każdej komplikacji modelu zweryfikować, czy wprowadzona zmiana w sposób istotny poprawiła jakość modelu. Tak więc, jeżeli stwierdzone zostanie, że realizacja któregoś z etapów nie poprawia, w wymaganym przez badacza stopniu, jakości modelu, należy ten etap pominąć i przejść do następnego (np. z etapu 1 do 3). Mogą również wystąpić sytuacje, w których wprowadzenie komplikacji dla

pewnych poziomów poprawi precyzję szacunku w przypadku niektórych zmiennych a w przypadku innych nie. W takim wypadku należy wprowadzić tylko te zmiany, dzięki którym uzyskano poprawę precyzji szacunku. Weryfikacja zasadności wprowadzania kolejnych komplikacji zgodnych z przedstawianą metodologią modelowania wielopoziomowego jest niezwykle istotna, ponieważ niektóre dane mogą w rzeczywistości nie mieć struktury dwupoziomowej.

Warunkiem koniecznym występowania dwupoziomowej struktury badanej zmiennej jest dwupoziomowa struktura badanej populacji. Oznacza to, że jednostki statystyczne muszą dzielić się na skończoną liczbę rozłącznych i pokrywających całą zbiorowość grup. Kolejnym warunkiem koniecznym jest występowanie zróżnicowania poziomu badanej zmiennej w różnych grupach. Zróżnicowanie to wynikać może z bezpośredniej zależności pomiędzy badaną zmienną a przynależnością jednostki badania do grupy. Klasycznym podawanym w literaturze przykładem takiej sytuacji jest zróżnicowanie poziomu nauczania wynikające zarówno z indywidualnych zdolności i predyspozycji ucznia (czynników na poziomie pierwszym – jednostkowym) oraz kwalifikacji nauczyciela oraz stosowanej metody nauczania (czynniki na poziomie drugim – grupowym) (por. Hox, 2002). Jako inną przyczynę zróżnicowania między grupami wskazać można zależności badanej zmiennej oraz podziału na grupy z pewną ukrytą, często niemierzalną, zmienną. Jako przykład podać można relację między aktywnością zawodową i stopniem rozwoju regionów. Przestrzenne zróżnicowanie czynników określających popyt na pracę, lokalizacja zasobów naturalnych, zakładów produkcyjnych, rozwój infrastruktury technicznej, komunikacyjnej, edukacyjnej są istotnymi determinantami określającymi aktywność ekonomiczną ludności obok czynników charakteryzujących przedsiębiorczość poszczególnych osób. Jeżeli struktura zmiennej objaśnianej spełnia obydwa warunki konieczne; występowanie dwupoziomowej struktury populacji oraz przesłanek aby podejrzewać, że poziom badanej zmiennej jest zróżnicowany pomiędzy grupami, należy jeszcze zweryfikować, czy występujące pomiędzy grupami zróżnicowanie jest istotne na obranym poziomie istotności. W tym celu zastosować można np. test analizy wariancji.

2. ALGORYTM KONSTRUKCJI MODELU DWUPOZIOMOWEGO

W opisie konstrukcji modelu dwupoziomowego przyjęto następujące oznaczenia:

n – liczebność całej próby,

J – liczba grup na pierwszym poziomie (liczba jednostek na drugim poziomie),

j – indeks obserwacji wskazującej na jej przynależność do *j-tej* grupy na pierwszym poziomie ($j = 1,\ldots,J$),

n_j – liczebność próby w *j-tej* grupie ($\sum_{j=1}^{J} n_j = n$),

i – indeks obserwacji wewnątrz *j-tej* grupy ($i=1,\ldots,n_j$),

Y_{ij} – wartość zmienna 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, $q = 1,\ldots,Q$.

W omawianej metodologii modelowania dwupoziomowego przyjmuje się następujące założenia. Badana zmienna Y ma rozkład normalny: $Y_{ij} \sim N(a_j; \sigma^2)$, *dla* $j=1,...,J$, $i=1,\ldots,n_j$. Należy interpretować to tak, że zakładamy równość wariancji zmiennej Y w całej populacji (co można zweryfikować przy pomocy testu Bartleta) oraz równość wartości oczekiwanej zmiennej Y w poszczególnych grupach.

Weryfikacji hipotezy o dwupoziomowej strukturze danych dokonuje się przy użyciu testu analizy wariancji (por. Krzyśko, 1996). Hipoteza zerowa tego testu głosi równość średnich badanej zmiennej policzonych we wszystkich J grupach na pierwszym poziomie. Hipoteza alternatywna głosi zaś, że z J grup na pierwszym poziomie można wybrać co najmniej dwie grupy, w których średnie policzone z wartości badanej zmiennej różnią się na zadanym poziomie istotności.

ETAP 0 Klasyczna regresja liniowa

Celem porównania w późniejszych etapach, wyznaczone zostały dwie funkcje regresji liniowej. Pierwsza funkcja, w której nie uwzględniono żadnych zmiennych objaśniających:

$$Y_{ij} = \gamma_{00}^{a} + r_{ij}^{a}, \quad r_{ij}^{a} \sim N\left(0; \sigma_a^2\right), \tag{1}$$

gdzie:

γ_{00}^{a} – estymator poziomu zmiennej objaśnianej z pierwszego poziomu, nie uwzględniający przynależności jednostek do grupy,

$r_{ij}^{a} \sim N\left(0;\sigma_a^2\right)$ – niezależne reszty dla jednostek z pierwszego poziomu,

σ_a^2 – wariancja reszt dla pierwszego poziomu.

W drugiej funkcji regresji jako zmienne objaśniające przyjęto zmienne z pierwszego poziomu:

$$Y_{ij} = \gamma_{00}^{b} + \sum_{p=1}^{P}\left(\gamma_{p0}^{b} X_{pij}\right) + r_{ij}^{b}, \quad r_{ij}^{b} \sim N\left(0; \sigma_b^2\right), \tag{2}$$

gdzie:

γ_{00}^{b} – wyraz wolny funkcji regresji dla jednostek pierwszego poziomu,

$r_{ij}^{b} \sim N\left(0; \sigma_b^2\right)$ – niezależne reszty dla jednostek z pierwszego poziomu, składnik resztowy,

σ^2 – wariancja reszt dla pierwszego poziomu,

γ_{p0}^{b} – współczynnik kierunkowy liniowej funkcji regresji dla *p-tej* zmiennej objaśniającej z pierwszego poziomu, $p = 1,...,P$.

W celu estymacji parametrów zamiast klasycznej metody najmniejszych kwadratów sugeruje się raczej metodę największej wiarogodności[1], co w dalszych etapach pozwala na porównanie jakości oszacowań otrzymanych przy pomocy różnych modeli.

ETAP 1 Model zerowy (null model)

W tym etapie zmienna Y_{ij} objaśniana jest wyłącznie przez przynależność jednostek z pierwszego poziomu do grup na drugim poziomie, bez użycia zmiennych objaśniających. Model można zapisać w wersji dwurównaniowej (dla każdego poziomu oddzielnie):

Na poziomie pierwszym model zapisać można przy pomocy równania:

$$Y_{ij} = \beta_{0j}^{I} + r_{ij}^{I}, \quad r_{ij}^{I} \sim N\left(0; \sigma_{I}^{2}\right), \tag{3}$$

gdzie:

β_{0j}^{I} – estymator punktowy zmiennej objaśnianej dla jednostek pierwszego poziomu należących do *j-tej* grupy,

$r_{ij}^{I} \sim N\left(0; \sigma_{I}^{2}\right)$ – niezależne reszty dla jednostek z pierwszego poziomu,

σ_{I}^{2} – wariancja reszt dla pierwszego poziomu,

Na drugim poziomie model jest następujący:

$$\beta_{0j}^{I} = \gamma_{00}^{I} + e_{0j}^{I}, \quad e_{0j}^{I} \sim N\left(0; \tau_{00}^{I}\right), \tag{4}$$

gdzie:

γ_{00}^{I} – estymator zmiennej objaśnianej z drugiego poziomu,

$e_{0j}^{I} \sim N\left(0; \tau_{00}^{I}\right)$ – niezależne reszty dla jednostek drugiego poziomu,

τ_{00}^{I} – wariancja reszt dla drugiego poziomu,

Ogólny model przyjmuje zatem postać (w wersji jednorównaniowej):

$$Y_{ij} = \gamma_{00}^{I} + e_{0j}^{I} + r_{ij}^{I}. \tag{5}$$

Porównanie wyników z liniową funkcją regresji bez zmiennych objaśniających pozwoli ocenić, czy samo uwzględnienie struktury dwupoziomowej poprawia precyzję szacunku.

ETAP 2 Model z losowym wyrazem wolnym (random intercept model)

W następnym etapie uwzględniony został wpływ przynależności każdej z jednostek z poziomu pierwszego (*i*) do danej jednostki z poziomu drugiego (*j*), jednak tylko w zakresie zróżnicowania wyrazu wolnego. Dopuszczona zostanie zatem możliwość

[1] W części opracowań naukowych używa się słowa wiarygodność; znaczenie słów wiarogodność i wiarygodność jest takie samo.

zróżnicowania wartości zmiennej objaśnianej z pierwszego poziomu dla różnych grup na poziomie drugim. Zakładamy jednak, że nachylenie krzywej regresji pozostanie stałe bez względu przynależność na poziomie drugim.

Model dla jednostek pierwszego poziomu jest następujący:

$$Y_{ij} = \beta_{0j}^{II} + \sum_{p=1}^{P} \left(\beta_{pj}^{II} X_{pij}\right) + r_{ij}^{II}, \quad r_{ij}^{II} \sim N\left(0; \sigma_{II}^2\right), \tag{6}$$

gdzie:

β_{0j}^{II} – wyraz wolny funkcji regresji dla jednostek pierwszego poziomu należących do j-tej grupy,

$r_{ij}^{II} \sim N\left(0; \sigma_{II}^2\right)$ – niezależne reszty dla jednostek z pierwszego poziomu,

σ_{II}^2 – wariancja reszt dla pierwszego poziomu,

β_{pj}^{II} – współczynnik kierunkowy liniowej funkcji regresji dla *p-tej* zmiennej objaśniającej z pierwszego poziomu, $p= 1,...,P$.

Dla jednostek drugiego poziomu model opisują następujące wzory:

$$\beta_{0j}^{II} = \gamma_{00}^{II} + e_{0j}^{II}, \quad e_{0j}^{II} \sim N\left(0; \tau_{00}^{II}\right), \tag{7}$$

$$\beta_{pj}^{II} = \gamma_{p0}^{II}, \text{ dla } p = 1, \ldots, P, \tag{8}$$

gdzie:

γ_{00}^{II} – estymator zmiennej objaśnianej z drugiego poziomu,

$e_{0j}^{II} \sim N\left(0; \tau_{00}^{II}\right)$ – niezależne reszty dla jednostek drugiego poziomu,

τ_{00}^{II} – wariancja reszt dla drugiego poziomu (wariancja jednostek drugiego poziomu),

γ_{p0}^{II} – współczynnik kierunkowy przy *p-tej* zmiennej z pierwszego poziomu, niezależny od jednostek drugiego poziomu, $p = 1,\ldots,P$.

W wersji jednorównaniowej model przyjmuje więc następującą postać:

$$Y_{ij} = \gamma_{00}^{II} + e_{0j}^{II} + \sum_{p=1}^{P} \left(\gamma_{p0}^{II} X_{pij}\right) + r_{ij}^{II}. \tag{9}$$

Tak wyznaczony model może być traktowany jako bezpośrednie rozwinięcie modelu opisanego w etapie 1, przez dodanie zmiennych objaśniających z pierwszego poziomu. Jednak może być traktowany również jako rozwinięcie modelu klasycznej regresji liniowej ze zmiennymi objaśniającymi z pierwszego poziomu, opisanego w etapie 0. W związku z powyższym należy wykazać jego wyższość nad oboma z nich.

ETAP 3 Model z losowym wyrazem wolnym zależnym od zmiennych z drugiego poziomu

W porównaniu z modelem z etapu drugiego, model wzbogacony zostanie o zmienne objaśniające z drugiego poziomu. Oznacza to, że wyraz wolny zależy nie tylko od przynależności do grup na drugim poziomie, ale objaśniany jest także przy pomocy zmiennych z drugiego poziomu.

Na pierwszym poziomie otrzymano zatem model opisany wzorem:

$$Y_{ij} = \beta_{0j}^{III} + \sum_{p=1}^{P} \left(\beta_{pj}^{III} X_{pij}\right) + r_{ij}^{III}, \quad r_{ij}^{III} \sim N\left(0; \sigma_{III}^2\right), \tag{10}$$

gdzie:

β_{0j}^{III} – wyraz wolny funkcji regresji dla jednostek pierwszego poziomu należących do *j-tej* grupy,

$r_{ij}^{III} \sim N\left(0; \sigma_{III}^2\right)$ – niezależne reszty dla jednostek z pierwszego poziomu,

σ_{III}^2 – wariancja reszt dla pierwszego poziomu,

β_{pj}^{III} – współczynnik kierunkowy liniowej funkcji regresji dla *p-tej* zmiennej objaśniającej z pierwszego poziomu, $p = 1,...,P$.

Na drugim poziomie model jest następujący:

$$\beta_{0j}^{III} = \gamma_{00}^{III} + \sum_{q=1}^{Q} \left(\gamma_{0q}^{III} Z_{qj}\right) + e_{0j}^{III}, \quad e_{0j}^{III} \sim N\left(0; \tau_{00}^{III}\right), \tag{11}$$

$$\beta_{pj}^{III} = \gamma_{p0}^{III}, \text{ dla } p = 1, \ldots, P, \tag{12}$$

gdzie:

γ_{00}^{III} – wyraz wolny funkcji regresji dla jednostek drugiego poziomu,

$e_{0j}^{III} \sim N\left(0; \tau_{00}^{III}\right)$ – niezależne reszty dla drugiego poziomu,

τ_{00}^{III} – wariancja reszt dla jednostek drugiego poziomu,

γ_{0q}^{III} – współczynnik kierunkowy liniowej funkcji regresji dla *q-tej* zmiennej z drugiego poziomu, $q = 1,\ldots,Q$,

γ_{p0}^{III} – współczynnik kierunkowy przy *p-tej* zmiennej z pierwszego poziomu, niezależny od jednostek drugiego poziomu, $p = 1,\ldots,P$.

Ogólny model w wersji jednorównaniowej można zatem zapisać następująco:

$$Y_{ij} = \gamma_{00}^{III} + \sum_{q=1}^{Q} \left(\gamma_{0q}^{III} Z_{qj}\right) + e_{0j}^{III} + \sum_{p=1}^{P} \left(\gamma_{p0}^{III} X_{pij}\right) + r_{ij}^{III}. \tag{13}$$

Model ten jest rozwinięciem modelu z losowym wyrazem wolnym i zmiennymi objaśniającymi z pierwszego poziomu (etap 2), przez dodanie do niego zmiennej objaśniającej z drugiego poziomu. Podobnie jak w poprzednich etapach należy zastosować procedurę weryfikującą zasadność jego zastosowania.

ETAP 4 Model z losowym współczynnikiem regresji

W tym etapie dopuszczona zostanie możliwość zróżnicowania grup z drugiego poziomu nie tylko ze względu na poziom szacowanej zmiennej, ale również ze względu na kształt zależności pomiędzy tą zmienną a zmiennymi objaśniającymi z pierwszego poziomu.

Tak skonstruowany model dla jednostek pierwszego poziomu opisany jest wzorem:

$$Y_{ij} = \beta_{0j}^{IV} + \sum_{p=1}^{P} \left(\beta_{pj}^{IV} X_{pij}\right) + r_{ij}^{IV}, \quad r_{ij}^{IV} \sim N\left(0; \sigma_{IV}^2\right), \tag{14}$$

gdzie:

β_{0j}^{IV} – wyraz wolny funkcji regresji dla jednostek pierwszego poziomu należących do j-tej grupy,

$r_{ij}^{IV} \sim N\left(0;\sigma_{IV}^2\right)$ – niezależne reszty dla jednostek z pierwszego poziomu,

σ_{IV}^2 – wariancja reszt dla pierwszego poziomu,

β_{pj}^{IV} – współczynnik kierunkowy liniowej funkcji regresji dla *p-tej* zmiennej objaśniającej z pierwszego poziomu, $p = 1,...,P$.

Z kolei na drugim poziomie model opisany jest wzorami:

$$\beta_{0j}^{IV} = \gamma_{00}^{IV} + \sum_{q=1}^{Q} \left(\gamma_{0q}^{IV} Z_{qj}\right) + e_{0j}^{IV}, \quad e_{0j}^{IV} \sim N\left(0; \tau_{00}^{IV}\right), \tag{15}$$

$$\beta_{pj}^{IV} = \gamma_{p0}^{IV} + e_{pj}^{IV}, \quad e_{pj}^{IV} \sim N\left(0; \tau_{pp}^{IV}\right), \text{ dla } p = 1, \ldots, P, \tag{16}$$

gdzie:

γ_{00}^{IV} – wyraz wolny funkcji regresji dla jednostek drugiego poziomu,

$e_{0j}^{IV} \sim N\left(0; \tau_{00}^{IV}\right)$ – niezależne reszty dla jednostek drugiego poziomu,

τ_{00}^{IV} – wariancja reszt dla drugiego poziomu,

γ_{0q}^{IV} – współczynnik kierunkowy liniowej funkcji regresji dla *q-tej* zmiennej z drugiego poziomu, $q = 1,\ldots,Q$,

γ_{p0}^{IV} – współczynnik kierunkowy dla *p-tej* zmiennej z pierwszego poziomu, niezależny od jednostek drugiego poziomu, $p = 1,\ldots,P$,

e_{pj}^{IV} – składnik resztowy z drugiego poziomu dla współczynnika kierunkowego przy *p-tej* zmiennej z pierwszego poziomu, $p = 1,\ldots,P$.

Zatem ogólnym model w wersji jednorównaniowej można zapisać w następujący sposób:

$$Y_{ij} = \gamma_{00}^{IV} + \sum_{q=1}^{Q} \left(\gamma_{0q}^{IV} Z_{qj}\right) + e_{0j}^{IV} + \sum_{p=1}^{P} \left(\left(\gamma_{p0}^{IV} + e_{pj}^{IV}\right) X_{pij}\right) + r_{ij}^{IV}. \tag{17}$$

Dla każdej zmiennej objaśniającej z pierwszego poziomu należy sprawdzić, czy uwzględnienie losowego charakteru jej współczynnika kierunkowego w sposób istotny poprawia

jakość modelu w porównaniu z modelem bez losowych współczynników kierunkowych (etap 3). W dalszych rozważaniach należy przyjąć model rozszerzony o losowość tylko tych z współczynników kierunkowych, w przypadku których uwzględnienie losowego charakteru poprawia jakość modelu.

ETAP 5 Model z losowym współczynnikiem regresji zależnym od zmiennych z drugiego poziomu

W modelu regresji liniowej dla losowego wyrazu wolnego oraz losowego współczynnika kierunkowego objaśnianego przy użyciu zmiennych objaśniających z drugiego poziomu uwzględnia się możliwość wpływu przynależności do jednostki drugiego poziomu tak na poziom jak i kształt zależności pomiędzy szacowaną zmienną a zmiennymi objaśniającymi z pierwszego poziomu. Dodatkowo objaśnia się zmienność szacowanej cechy na poziomie pierwszym przy pomocy zmiennych określonych na poziomie drugim. Tak więc przyjmuje się, że współczynniki kierunkowe dla zmiennych objaśniających z pierwszego poziomu nie tylko różnią się zależnie od przynależności do jednostek z drugiego poziomu, ale są objaśniane przez zmienne z drugiego poziomu.

Model dla jednostek pierwszego poziomu jest następujący:

$$Y_{ij} = \beta_{0j}^{V} + \sum_{p=1}^{P} \left(\beta_{pj}^{V} X_{pij}\right) + r_{ij}^{V}, \quad r_{ij}^{V} \sim N\left(0; \sigma_{V}^{2}\right), \tag{18}$$

gdzie:

β_{0j}^{V} – wyraz wolny funkcji regresji dla jednostek pierwszego poziomu należących do *j-tej* grupy,

$r_{ij}^{V} \sim N\left(0; \sigma_{V}^{2}\right)$ – niezależne reszty dla jednostek z pierwszego poziomu, σ_{V}^{2} – wariancja reszt dla pierwszego poziomu,

β_{pj}^{V} – współczynnik kierunkowy liniowej funkcji regresji dla *p-tej* zmiennej objaśniającej z pierwszego poziomu, $p = 1,...,P$.

Na drugim poziomie model jest następujący:

$$\beta_{0j}^{V} = \gamma_{00}^{V} + \sum_{q=1}^{Q} \left(\gamma_{0q}^{V} Z_{qj}\right) + e_{0j}^{V}, \quad e_{0j}^{V} \sim N\left(0; \tau_{00}^{V}\right), \tag{19}$$

$$\beta_{pj}^{V} = \gamma_{p0}^{V} + \sum_{q=1}^{Q} \left(\gamma_{pq}^{V} Z_{qj}\right) + e_{pj}^{V}, \quad e_{pj}^{V} \sim N\left(0; \tau_{pp}^{V}\right), \text{ dla } p = 1, \ldots, P, \tag{20}$$

gdzie:

γ_{00}^{V} – wyraz wolny funkcji regresji dla jednostek drugiego poziomu,

$e_{0j}^{V} \sim N\left(0; \tau_{00}^{V}\right)$ – niezależne reszty dla jednostek drugiego poziomu,

τ_{00}^{V} – wariancja reszt dla drugiego poziomu,

γ_{0q}^{V} – współczynnik kierunkowy liniowej funkcji regresji dla *q-tej* zmiennej z drugiego poziomu, $q = 1,\ldots,Q$,

γ_{p0}^{V} – wyraz wolny liniowej funkcji regresji objaśniającej wartość współczynników kierunkowych z pierwszego poziomu dla *p-tej* zmiennej, $p = 1,\ldots,P$,

γ_{pq}^{V} – współczynnik kierunkowy *q-tej* zmiennej z drugiego poziomu w liniowej funkcji regresji objaśniającej wartość współczynnika kierunkowego *p-tej* zmiennej z pierwszego poziomu, $p = 1,\ldots,P$, $q = 1,\ldots,Q$,

e_{pj}^{V} – składnik resztowy z drugiego poziomu we współczynniku kierunkowym dla *p-tej* zmiennej z pierwszego poziomu, $p = 1,\ldots,P$,

τ_{pp}^{V} – wariancja składnika resztowego z drugiego poziomu we współczynniku kierunkowym dla *p-tej* zmiennej z pierwszego poziomu, $p = 1,\ldots,P$.

W wersji jednorównaniowej model przyjmuje postać:

$$Y_{ij} = \gamma_{00}^{V} + \sum_{q=1}^{Q}\left(\gamma_{0q}^{V} Z_{qj}\right) + e_{0j}^{V} + \sum_{p=1}^{P}\left(\left(\gamma_{p0}^{V} + \sum_{q=1}^{Q}\left(\gamma_{pq}^{V} Z_{qj}\right) + e_{pj}^{V}\right) X_{pij}\right) + r_{ij}^{V}. \tag{21}$$

W tym etapie należy pamiętać, że nie każda zmienna z pierwszego poziomu nadaje się do objaśniania wybranej zmiennej z poziomu drugiego. Sytuacja taka uwarunkowana może być tak względami merytorycznymi, jak również brakiem poprawy jakości modelu. W takim wypadku wartość odpowiedniego współczynnika kierunkowego jest równa zero ($\gamma_{pq}^{V} = 0$), tym samym *q-ta* zmienna z drugiego poziomu nie jest używana do konstrukcji współczynnika kierunkowego przy *p-tej* zmiennej z poziomu pierwszego.

3. KRYTERIA OCENY JAKOŚCI DOPASOWANIA MODELU DWUPOZIOMOWEGO

3.1 FUNKCJA WIAROGODNOŚCI L

Jednym z kryteriów oceny dopasowania modelu, oraz podstawą wielu innych kryteriów, jest maksimum opisywanej wzorem (26) funkcji wiarogodności (por. Harville, 1974).

$$L\left(Y\left|\Theta,\sigma^2\right.\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} * exp\left\{-\frac{1}{2} * \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_{ij}\right)^2}{\sigma^2}\right\}, \tag{22}$$

gdzie:

Θ – zbiór parametrów szacowanych w modelu (wsp. kierunkowe, wyraz wolny itp.),

σ – odchylenie standardowe składnika losowego modelu, $r_{ij} \sim N\left(0;\sigma^2\right)$,

Y_{ij} – rzeczywista wartość zmiennej objaśnianej,

$\hat{Y}_{ij}$ – oszacowanie zmiennej objaśnianej przy pomocy rozważanej funkcji regresji.

Współczynniki funkcji regresji tworzonej zgodnie z metodą największej wiarogodności dobierane są tak, aby maksymalizować wartość funkcji wiarogodności (por. wzór (23)). Warunek tan można zapisać:

$$\bigwedge_{j=1,\ldots,J} \bigwedge_{i=1,\ldots,n_j} L\left(Y \left| \hat{\Theta}, \sigma^2 \right.\right) = \sup_{\Theta \epsilon R^k} L\left(Y \left| \Theta, \sigma^2 \right.\right), \tag{23}$$

gdzie:

$\hat{\Theta}$ – oszacowanie współczynników modelu wyznaczone przy pomocy metody największej wiarogodności,

k – liczba współczynników estymowanych w modelu.

Zazwyczaj łatwiej jest znaleźć supremum logarytmu naturalnego z funkcji wiarogodności niż supremum tej funkcji. Ponieważ logarytm naturalny jest funkcją rosnącą w całej dziedzinie, oszacowania zbirów szacowanych parametrów są identyczne. Zatem problem sprowadzić można do znalezienia takiego oszacowania $\hat{\Theta}$, aby spełniony był warunek:

$$\bigwedge_{j=1,\ldots,J} \bigwedge_{i=1,\ldots,n_j} lnL\left(Y \left| \hat{\Theta}, \sigma^2 \right.\right) = \sup_{\Theta \epsilon \mathrm{R}^k} lnL\left(Y \left| \Theta, \sigma^2 \right.\right). \tag{24}$$

Stąd też wykorzystywana funkcja przyjmuje postać:

$$lnL\left(Y \left| \Theta, \sigma^2 \right.\right) = -\frac{n}{2} * ln\left(2\pi\right) - n * ln\left(\sigma\right) - \frac{1}{2} * \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \hat{Y}_{ij}\right)^2}{\sigma^2}. \tag{25}$$

Można wykazać, że tak wyznaczone oszacowanie $\hat{\Theta}$ jest identyczne z oszacowaniem wyznaczonym przy pomocy klasycznej metody najmniejszych kwadratów (por. Rydlewski, 2009). Na podstawie wartości supremum $lnL\left(Y \left| \hat{\Theta}, \sigma^2 \right.\right)$ można wyznaczyć dodatkowo kilka współczynników świadczących o jakości dopasowania rozpatrywanego modelu do danych rzeczywistych. Część z nich zostanie przedstawiona poniżej.

3.2 KRYTERIUM INFORMACYJNE AKAIKE'A

Pierwszym z kryteriów opartych na funkcji wiarogodności jest kryterium informacyjne Akaike'a (AIC) (por. wzór (26)).

$$AIC = -2 * ln\left(L\right) + 2 * p, \tag{26}$$

gdzie:

L – maksymalna wartość funkcji wiarogodności modelu,

p – liczba szacowanych w modelu parametrów.

Wysoka wartość funkcji wiarogodności informuje o dobrym dopasowaniu modelu. Ponieważ nadmierny wzrost liczby zmiennych objaśniających uznawany jest za niekorzystny, kryterium AIC uwzględnia fakt, iż zbyt duża liczba szacowanych parametrów

obniża wartość modelu. Model z minimalną wartością AIC jest uznawany, według tego kryterium, za najlepiej dopasowany do danych (por. Sakamoto, Ishiguro, Kitagawa, 1986).

3.3 BAYESOWSKIE KRYTERIUM INFORMACYJNE

Kolejnym kryterium pozwalającym ocenić jakość dopasowania modelu jest Bayesowskie kryterium informacyjne (BIC). Wartość jaką przyjmuje współczynnik BIC wyznacza się zgodnie ze wzorem (31) (por. Schwarz, 1978).

$$BIC = -2 * ln(L) + ln(n) * p, \tag{27}$$

gdzie:

L – maksymalna wartość funkcji wiarogodności modelu,

n – liczba obserwacji w próbie (z pierwszego poziomu),

p – liczba szacowanych w modelu parametrów.

Porównując kryterium BIC z AIC, można stwierdzić, że podobnie jak poprzednio uwzględnia ono dodatni wpływ wysokiej wartości funkcji wiarogodności modelu oraz ujemne oddziaływanie zbyt dużej liczby szacowanych parametrów. Jednak znaczenie liczby szacowanych parametrów uzależniono od liczebności próby. Uznano je za istotniejsze, gdy próba jest liczna, a za mniej istotne, gdy próba jest mała. Ponieważ $ln(74) \approx 2$, dla sytuacji gdy liczebność próby jest większa bądź równa 8, kryterium BIC surowiej „karze" model za zwiększoną liczbę szacowanych parametrów niż AIC. Oczywiście, tak jak w przypadku kryterium AIC, za najlepszy uznawany jest model o najmniejszym współczynniku BIC. Kryterium BIC jest znane również jako SBC (Kryterium Bayesowskie Schwarza).

3.4 TEST ILORAZU WIAROGODNOŚCI χ^2

Przy pomocy testu χ^2 porównuje się jakość dopasowania dwóch różnych modeli: A oraz jego rozszerzenia B. O modelach tych zakładamy, że B, jako rozszerzenie A, jest lepiej dopasowany do danych empirycznych. Weryfikacja tej hipotezy sprowadza się do sprawdzenia, czy różnica w jakości obu modeli jest statystycznie istotna (por. Lin, 1997; Goldstein, 2003). Tak więc hipoteza zerowa zakłada, że model B nie jest istotnie lepszy od modelu A. Natomiast hipoteza alternatywna głosi, że wiarogodność modelu B jest statystycznie istotnie wyższa niż wiarogodność modelu A (por. wzór (28)). Odrzucenie hipotezy zerowej świadczy o tym, że warto rozszerzyć model A do modelu B i przyjąć model B za obowiązujący w dalszych analizach, w przypadku braku podstaw do odrzucenia hipotezy zerowej nie należy wprowadzać rozważanej zmiany do modelu A i pozostawić go obowiązującym w dalszych rozważaniach. Układ hipotez:

$$\begin{cases} H_0: \sigma_A = \sigma_B \\ H_1: \sigma_A > \sigma_B \end{cases}, \tag{28}$$

gdzie:

σ_A – odchylenie standardowe składnika losowego modelu A,

σ_B – odchylenie standardowe składnika losowego modelu B.

W celu weryfikacji hipotezy zerowej stosuje się statystykę testową następującej postaci:

$$\chi^2 = 2 * lnL_B - 2 * lnL_A = 2n * ln\left(\frac{\sigma_A}{\sigma_B}\right) + \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_{ij}^{A}\right)^2}{\sigma_A^2} - \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_{ij}^{B}\right)^2}{\sigma_B^2}, \quad (29)$$

gdzie:

L_A – supremum funkcji wiarogodności w modelu A,

L_B – supremum funkcji wiarogodności w modelu B,

$\hat{Y}_{ij}^{A}$ – oszacowanie zmiennej objaśnianej przy pomocy modelu A,

$\hat{Y}_{ij}^{B}$ – oszacowanie zmiennej objaśnianej przy pomocy modelu B.

Przy prawdziwości hipotezy zerowej statystyka χ^2 opisana jest więc wzorem:

$$\chi^2\big|_{H_0} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[\left(Y_{ij} - \hat{Y}_{ij}^{A}\right)^2 - \left(Y_{ij} - \hat{Y}_{ij}^{B}\right)^2\right]}{\sigma_A^2}. \quad (30)$$

Można wykazać, że tak wyznaczona statystyka przy prawdziwości hipotezy zerowej ma rozkład χ^2 z liczbą stopni swobody obliczaną jako różnica pomiędzy liczbą parametrów szacowanych w modelu B a liczbą parametrów szacowaną w modelu A:

$$\chi^2\big|_{H_0} \sim \chi^2\left(p_B - p_A\right), \quad (31)$$

gdzie:

p_A – liczba parametrów szacowanych w modelu A,

p_B – liczba parametrów szacowanych w modelu B.

Obszar krytyczny służący do weryfikacji powyższej hipotezy zerowej jest następującej postaci:

$$B = \left\{\underline{Y}: \chi^2 = 2 * lnL_B - 2 * lnL_A > \chi^2\left(1 - \alpha; p_B - p_A\right)\right\}, \quad (32)$$

gdzie:

α – przyjęty poziom istotności,

$\chi^2(a; b)$ – wartość kwantyla rozkładu χ^2 z b stopniami swobody z prawdopodobieństwa a.

W sytuacji gdy statystyka testowa przyjmuje wartość należącą do obszaru krytycznego hipotezę zerową odrzucamy na korzyść hipotezy alternatywnej. Oznacza to iż model B charakteryzuje się lepszym od modelu A dopasowaniem do danych empirycznych.

4. EMPIRYCZNA WERYFIKACJA PRZYDATNOŚCI MODELU DWUPOZIOMOWEGO NA PRZYKŁADZIE ESTYMACJI LICZBY PRACUJĄCYCH

Celem tej części artykułu jest oszacowanie liczby osób pracujących w przekroju powiatów. Z uwagi na zróżnicowanie badanych jednostek terytorialnych pod względem wielkości mierzonej liczbą ludności, w badaniach posłużono się wielkościami względnymi szacując udział pracujących wśród ludności w wieku produkcyjnym – zmienna objaśniana (Y_{ij}). Przyjęto założenie, iż tak mierzona aktywność ekonomiczna ludności w przekroju terytorialnym jest zdeterminowana, między innymi, poziomem rozwoju gospodarczego regionu utożsamianego z województwem. Poprawniejsze merytorycznie wydaje się przyjęcie założenia o zależności zróżnicowania aktywności ekonomicznej ludności zarówno od indywidualnych predyspozycji jednostek – osób (poziom pierwszy) oraz rozwoju gospodarczego regionu (poziom drugi). Badania takie wymagają jednak dostępności danych jednostkowych, np. z BAEL. Próba taka podjęta zostanie w kolejnym artykule. Podjęto zatem próbę konstrukcji modelu dwupoziomowego objaśniającego udział pracujących (Y_{ij}) w przekroju terytorialnym z jednostką pierwszego poziomu zdefiniowaną jako powiat i grupami, czyli inaczej poziomem drugim, w postaci województw. W celu uzyskania oszacowania liczby osób pracujących w powiatach, po zakończeniu obliczeń, wystarczy zmienną objaśnianą przemnożyć przez liczbę osób w wieku produkcyjnym.

4.1 ZMIENNE OBJAŚNIAJĄCE

Jako zmienne objaśniające na poziomie powiatów przyjęto stosunek salda dojazdów do pracy[2] do liczby mieszkańców w wieku produkcyjnym (X_{1ij}) oraz stosunek bezrobotnych do ludności w wieku produkcyjnym (X_{2ij}). Jako zmienną objaśniającą na poziomie województw przyjęto stosunek pracujących do ludności w wieku produkcyjnym (Z_{1j}). W celu zweryfikowania doboru zmiennych objaśniających obliczono współczynniki korelacji liniowej Pearsona każdej z tych zmiennych ze zmienną objaśnianą (por. tab. 1). Współczynniki te dla obu zmiennych objaśniających z poziomu pierwszego wskazują na występowanie zależności ze zmienną objaśnianą, na poziomie istotności 0,01. Współczynnik korelacji liniowej pomiędzy zmienną objaśnianą a zmienną objaśniającą z drugiego poziomu wskazuje również na występowanie zależności pomiędzy tymi zmiennymi na poziomie ufności zaledwie 0,01. Należy jednak zaznaczyć, że zmienna objaśniana określona jest dla zbiorowości powiatów, zaś zmienna objaśniająca z drugiego poziomu dla zbiorowości województw. W związku z tym, w celu policzenia współczynnika korelacji liniowej Pearsona konieczne było zintegrowanie obu tych zbiorowości. Zmienną z poziomu województw przeistoczono sztucznie w zmienną określoną na poziomie powiatów przypisując każdemu z nich wartość tej cechy odpowiadającą województwu do którego dany powiat należał. Autorka świadoma jest niedoskonałości takiego podejścia i w kolejnych pracach zamierza

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

zaprezentować inny sposób badania korelacji pomiędzy zmiennymi zdefiniowanymi na różnych poziomach.

Dane na temat dojazdów do pracy zaczerpnięto z opublikowanego przez Urząd Statystyczny w Poznaniu badania na temat dojazdów do pracy za rok 2006, opartego na danych pozyskanych z zasobów rejestrów podatkowych Ministerstwa Finansów. Wykorzystanie tego źródła zdeterminowało ustanowienie badania na rok 2006.

Tabela 1.
Współczynniki korelacji liniowej Pearsona pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi

Zmienna objaśniana	Zmienna objaśniające		
	X_1	X_2	Z
Y	-0,775104	-0,427117	0,169644
p-value	$<2{,}2e^{-16}$	$<2{,}2e^{-16}$	0,0009734

Źródło: Opracowanie własne na podstawie danych BDL

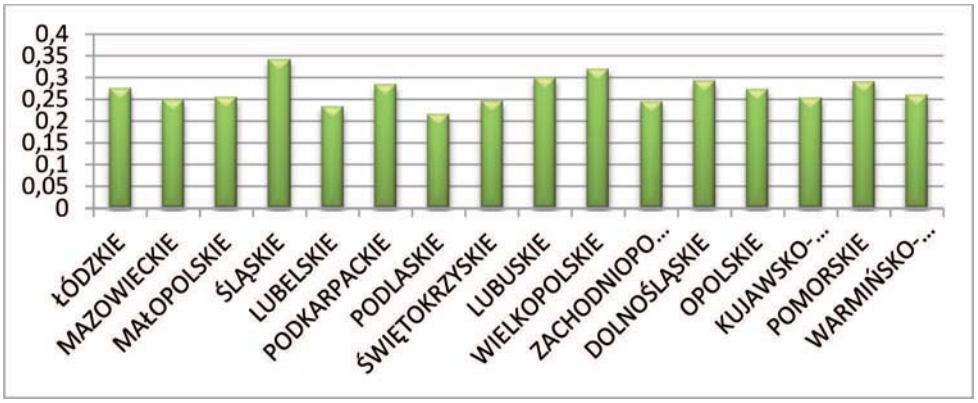
W badaniu uwzględniono wszystkie etapy komplikacji modelu opisane w części teoretycznej wraz z weryfikacją poprawy dopasowania modelu po każdym kroku. Obliczenia wykonane zostały przy pomocy biblioteki nlme[3] pakietu R 2.10.1. (por. Pinheiro, Bates, 2000; Bliese, 2012).

4.2 WERYFIKACJA HIPOTEZY O DWUPOZIOMOWEJ STRUKTURZE DANYCH, INACZEJ TEST ANALIZY WARIANCJI

Rozważania rozpoczęte zostaną od weryfikacji hipotezy, czy średni poziom zmiennej objaśnianej stosunek liczby pracujących do ludności w wieku produkcyjnym w powiatach różni się na poziomie istotności 0,05 pomiędzy województwami.

Jak zauważono wyżej, szacowana zmienna określająca stosunek pracujących do ludności w wieku produkcyjnym zalicza się do zmiennych opisujących aktywność zawodową, która jest zależna od stopnia rozwoju, który z kolei różni się w przekroju województw. W związku z tym również średni poziom badanej zmiennej w grupach powiatów należących do różnych województw różni się znacznie pomiędzy województwami. Najniższy średni poziom badanej zmienne odnotowano w powiatach słabo rozwiniętego województwa podlaskiego, wynosił on 0,21, co znaczy, że zatrudnionych było tam około 20 procent osób w wieku produkcyjnym. Z kolei w powiatach województwa śląskiego średni poziom stosunku zatrudnionych do ludności w wieku produkcyjnym wynosił aż 0,43, co oznacza, że zatrudnionych było tam więcej niż jedna trzecia osób w wieku produkcyjnym (por. rys. 1). Statystyczna istotność omówionego zróżnicowania poziomu średniego poziomu badanej zmiennej w powiatach różnych województw została zweryfikowana przy pomocy testu analizy wariancji. Na jego podstawie można na poziomie istotności 0,05 twierdzić, że występuje zróżnicowa-

[3] Nonlinear Mixed-Effects Models.



Rysunek 1. Średnie wartości badanej zmiennej wśród powiatów liczone w grupach (województwach)
Źródło: Opracowanie własne na podstawie BDL

nie województw ze względu na średnią wartość szacowanej zmiennej stosunek liczby pracujących do liczby osób w wieku produkcyjnym. Upoważnia to do przyjęcia założenia o dwupoziomowej strukturze danych oraz do konstrukcji modelu dwupoziomowego objaśniającego wartości zmiennej Y_{ij}.

4.3 KONSTRUKCJA MODELU

ETAP 0 Celem porównania w późniejszych etapach, wyznaczone zostały dwie funkcje regresji liniowej, pierwsza w której nie uwzględniono żadnych zmiennych objaśniających (dalej nazywana etap 0a):

$$\underset{(0{,}0054)}{Y_{ij} = 0,273812} + r^a_{ij}, \quad r^a_{ij} \sim N\left(0; 0,0108366\right).$$

oraz druga, w której jako zmienne objaśniające przyjęto dwie zmienne z pierwszego poziomu (dalej nazywana etap 0b):

$$\underset{(0,0087)}{Y_{ij} = 0,388495} - \underset{(0,0473)}{1,181261} * X_{1ij} - \underset{(0,0008)}{0,008116} * X_{2ij} + r^b_{ij}, \quad r^b_{ij} \sim N\left(0; 0,003307\right).$$

Dla obu modeli obliczone zostały wybrane miary jakości dopasowania (por. tab. 2), które będą punktem wyjścia do porównań w następnych etapach.

ETAP 1 W tym etapie zmienna Y_{ij} objaśniana jest wyłącznie przez przynależność powiatów do województw, bez użycia zmiennych objaśniających. Ogólny model w wersji jednorównaniowej przyjmuje postać:

$$\underset{(0,0087)}{Y_{ij} = 0,271122} + e^I_{0j} + r^I_{ij}, \quad r^I_{ij} \sim N\left(0; 0,010033\right), \quad e^I_{0j} \sim N\left(0; 0,000766\right).$$

Porównanie wyników z liniową funkcją regresji bez zmiennych objaśniających pozwoli ocenić, czy samo uwzględnienie struktury dwupoziomowej poprawi precyzję szacunku.

Tabela 2.

Wybrane kryteria oceny funkcji regresji liniowej

Etap	*Kryterium oceny*		
	lnL	*AIC*	*BIC*
etap 0a	313,3386	-622,6771	-614,8233
etap 0b	528,4533	-1048,907	-1033,220

Źródło: Opracowanie własne

Na podstawie otrzymanych wyników (por. tab. 3.), w szczególności niskiego prawdopodobieństwa popełnienia błędu pierwszego rodzaju dla statystyki χ^2, oceniającej poprawę wiarogodności modelu w stosunku do modelu uboższego o losowy wyraz wolny, można stwierdzić, że nastąpiła statystycznie istotna poprawa jakości modelu. Również wybrane kryteria oceny poprawy jakości modelu wskazują na jego wyższość nad modelem zwykłej regresji liniowej (por. tab. 2 i 3).

Tabela 3.

Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 1 oraz jego porównanie z etapem 0a

Etap	*Kryterium oceny*			*vs etap 0a*	
	AIC	*BIC*	*lnL*	*Chi2*	*p-value*
etap 1	-634,5586	-622,7779	320,2793	13,88151	0,0002

Źródło: Opracowanie własne na podstawie BDL

ETAP 2 W następnym etapie uwzględniony został wpływ przynależności każdego z powiatów (i) do danego województwa (j), jednak tylko w zakresie wyrazu wolnego. Oznacza to, że dopuszcza się możliwość zróżnicowania powiatów w przekroju województw ze względu na liczbę osób pracujących, jednak nachylenie krzywej regresji pozostanie stałe dla wszystkich województw.

W postaci ogólnej model opisany jest równaniem:

$$Y_{ij} = \underset{(0,0117)}{0,380805} - \underset{(0,0419)}{1,195488} * X_{1ij} - \underset{(0,0008)}{0,007748} * X_{2ij} + e_{0j}^{II} + r_{ij}^{II},$$

$$r_{ij}^{II} \sim N(0; 0,002519), \quad e_{0j}^{II} \sim N(0; 0,000886).$$

Tak wyznaczony model może być traktowany jako bezpośrednie rozwinięcie modelu opisanego w etapie 1. Polega ono na dodaniu zmiennych objaśniających z pierwszego poziomu. Model powyższy jest również rozwinięciem zwykłej regresji liniowej z dwoma zmiennymi objaśniającymi, opisanej w etapie 0b. Z tego względu zastosowano procedurę oceniającą poprawę jakości dopasowania modelu według przyjętych kryteriów w stosunku do obu z nich (por. tab. 4). Bardzo niskie prawdopodobieństwa

popełnienia błędu pierwszego rodzaju w obu testach wskazują na zdecydowaną poprawę jakości modelu z etapu 2 w stosunku do obu modeli prostszych. Zatem wyznaczony w etapie drugim model będzie punktem wyjścia w kolejnym kroku. Pozostałe kryteria oceny jakości dopasowania modelu również wskazują na wyraźną przewagę modelu opisanego w tym etapie zarówno nad modelem z etapu 0b jak i modelem z etapu 1 (por. tab. 4 z 2 i 3).

Tabela 4.

Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 2 oraz jego porównanie z etapem 0b oraz 1

Etap	***Kryterium oceny***			***vs etap 0b***		***vs etap 1***	
	AIC	***BIC***	***lnL***	***Chi2***	***p-value***	***Chi2***	***p-value***
etap 2	-1116,130	-1096,522	563,0648	69,22312	<,0001	485,571	<,0001

Źródło: Opracowanie własne

ETAP 3 W porównaniu z modelem z etapu drugiego, model wzbogacony został o zmienną objaśniającą stosunek liczby osób pracujących do liczby osób w wieku produkcyjnym z poziomu województw. W wersji jednorównaniowej zapisać można go jako:

$$Y_{ij} = \underset{(0,0524)}{0,218156} + \underset{(0,1616)}{0,511914} * Z_{1j} - \underset{(0,0418)}{1,201132} * X_{1ij} - \underset{(0,0008)}{0,007507} * X_{2ij} + e_{0j}^{III} + r_{ij}^{III},$$

$$r_{ij}^{III} \sim N(0; 0,002518), \quad e_{0j}^{III} \sim N(0; 0,000509).$$

Model ten jest rozwinięciem modelu z losowym wyrazem wolnym i zmiennymi objaśniającymi z poziomu powiatów (etap 2), przez dodanie zmiennej objaśniającej z poziomu województw. Zatem aby wprowadzenie go uznać za zasadne, należy sprawdzić, czy jest lepszy od tego modelu. Ponieważ na poziomie istotności 0,0115 możemy twierdzić, że wiarogodność modelu ze zmienną objaśniającą z drugiego poziomu jest wyższa od wiarogodności modelu bez tej zmiennej (por. tab. 5), model ten uznajemy za najlepszy z modeli dotąd rozważanych.

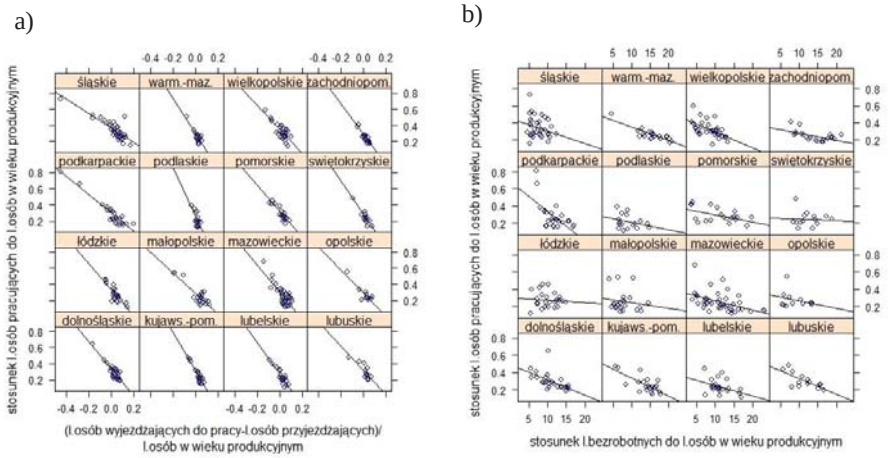
Tabela 5.

Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 3 oraz jego porównanie z etapem 2

Etap	***Kryterium oceny***			***vs etap 2***	
	AIC	***BIC***	***lnL***	***Chi2***	***p-value***
etap 3	-1120,524	-1097,01	566,2618	6,393965	0,0115

Źródło: Opracowanie własne

ETAP 4 W kolejnym etapie dopuszczona zostanie możliwość zróżnicowania województw nie tylko ze względu na poziom szacowanej zmiennej, ale również ze względu na kształt zależności pomiędzy tą zmienną a zmiennymi objaśniającymi X_{1ij} oraz X_{2ij}.



Rysunek 2. a) Linie regresji liniowej dla zmiennej X_1; b) Linie regresji liniowej dla zmiennej X_2
Źródło: Opracowanie własne

W celu graficznej weryfikacji, dla której ze zmiennych objaśniających X_{1ij} oraz X_{2ij} współczynnik kierunkowy zależy od przynależności powiatów do województwa wykreślono linie regresji dla obu tych zmiennych i dla każdego z województwa osobno. Na podstawie tych wykresów zaobserwowano różnice w nachyleniach linii regresji dla obu zmiennych. Większe różnice stwierdzono dla zmiennej stosunek liczby bezrobotnych do liczby osób w wieku produkcyjnym (por. rys. 2) Ponieważ jednak uwzględnienie losowego charakteru współczynnika kierunkowego zmiennej X_{2ij} nie poprawiło w sposób istotny jakości modelu w porównaniu z modelem bez losowych współczynników kierunkowych (por. tab. 6) zdecydowano nie uwzględniać losowego charakteru tego współczynnika kierunkowego. Następnie sprawdzono, jak zmieni się precyzja szacunku dla modelu z losowym współczynnikiem kierunkowym dla zmiennej X_{1ij}. Ponieważ uwzględnienie losowego charakteru współczynnika kierunkowego przy X_{1ij} spowodowało istotną poprawę jakości modelu (por. tab. 7), w dalszych rozważaniach przyjęto model rozszerzony w ten sposób.

Tabela 6.
Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 4, dla ulosowionego współczynnika kierunkowego zmiennej X_{2ij} oraz jego porównanie z etapem 3

Etap	***Kryterium oceny***			***vs etap 3***	
	AIC	***BIC***	***lnL***	***Chi2***	***p-value***
Losowy współczynnik kierunkowy dla X_2	-1119,554	-1088,202	567,7768	3,029897	0,2198

Źródło: Opracowanie własne

Tabela 7.

Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 4, dla ulosowionego współczynnika kierunkowego zmiennej X_{1ij} oraz jego porównanie z etapem 3

Etap	***Kryterium oceny***			***vs etap 3***	
	AIC	***BIC***	***lnL***	***Chi2***	***p-value***
Losowy współczynnik kierunkowy dla X_1	-1153,599	-1122,247	584,7993	37,07495	<,0001

Źródło: Opracowanie własne

Tak skonstruowany model w wersji jednorównaniowej ma postać:

$$Y_{ij} = \underset{(0,0464)}{0,2286} + \underset{(0,1426)}{0,4748} * Z_{1j} + e_{0j}^{IV} + \underbrace{\left(-1,43673 + e_{1j}^{IV}\right)}_{(0,0969)} * X_{1ij} - \underset{(0,0008)}{0,00693} * X_{pij} + r_{ij}^{IV},$$

$$r_{ij}^{IV} \sim N(0; 0,002183), \quad e_{0j}^{IV} \sim N(0; 0,000366), \quad e_{1j}^{IV} \sim N(0; 0,09209).$$

ETAP 5 W modelu regresji liniowej dla losowego wyrazu wolnego oraz losowego nachylenia z dodanymi zmiennymi objaśniającymi z poziomu województw dopuszczona jest nie tylko możliwość wpływu przynależności powiatu do województwa na poziom szacowanej zmiennej oraz kształt zależności pomiędzy nią a zmiennymi objaśniającymi na poziomie powiatów, ale także objaśniania szacowanej zmiennej na poziomie powiatów przy pomocy zmiennej na poziomie województw. Tak więc przyjęto, że współczynnik kierunkowy dla zmiennej X_{1ij} nie tylko różni się zależnie od przynależności powiatu do województwa, ale jest objaśniany przez zmienną Z_{1j} z poziomu województw. Zmiana ta spowodowała zwiększenie wiarogodności modelu. Wynik testu χ^2 pozwala poprawę tą uznać ze istotną na minimalnym poziomie istotności 0,0209, czyli na najczęściej przyjmowanym poziomie istotności 0,05 można stwierdzić poprawę. Kryterium AIC również świadczy o poprawie jakości modelu, jednak kryterium BIC, surowiej „karzące" za szacowanie dodatkowych parametrów, model ten klasyfikuje jako gorszy od poprzedniego (por. tab. 7 i 8). Po rozważeniu powyższych wyników zdecydowano model z etapu piątego przyjąć jako końcowy.

Tabela 8.

Wybrane kryteria oceny jakości dopasowania do danych modelu z etapu 5 oraz jego porównanie z etapem 4

Etap	***Kryterium oceny***			***vs etap 4***	
	AIC	***BIC***	***lnL***	***Chi2***	***p-value***
etap 5	-1156,934	-1121,688	587,4671	5,335611	0,0209

Źródło: Opracowanie własne

Opisany powyżej model w wersji jednorównaniowej zapisany może być równaniem:

$$Y_{ij} = \underset{(0,0463)}{0,22} + \underset{(0,1422)}{0,501} * Z_{1j} + e_{0j}^{V} + \left(\underset{(0,8187)}{-2,488} + \underset{(2,5433)}{3,278} * Z_{1j} + e_{1j}^{V}\right) * X_{1ij} - \underset{(0,0008)}{0,007} * X_{2ij} + r_{ij}^{V},$$

$$r_{ij}^{V} \sim N(0; 0,002175), \quad e_{0j}^{V} \sim N(0; 0,000358), \quad e_{1j}^{V} \sim N(0; 0,10136).$$

4.4 CAŁKOWITA POPRAWA DOPASOWANIA MODELU UZYSKANA DZIĘKI METODOLOGII MODELOWANIA DWUPOZIOMOWEGO

W tym akapicie przedstawiono poprawę jakości szacunku uzyskaną dzięki zastosowaniu funkcji regresji z modelu opisanego w etapie piątym w stosunku do zwykłej regresji liniowej ze zmiennymi objaśniającymi (etap 0b).

Tabela 9.
Wybrane kryteria oceny jakości dopasowania do danych modeli z etapów 0b i 5 oraz ich porównanie

Etap	***Kryterium oceny***			***Etap 5 vs etap 0b***	
	AIC	***BIC***	***lnL***		
etap 0b	-1048,907	-1033,22	528,4533	***Chi2***	***p-value***
etap 5	-1156,934	-1121,688	587,4671	118,0276	<,0001

Źródło: Opracowanie własne

Przeprowadzenie ostatniego testu χ^2 nie jest konieczne, ponieważ uzyskanie stopniowej poprawy dopasowania modelu w każdym z etapów jest gwarancją uzyskania poprawy całkowitej, ponieważ poprawa jakości modelu jest relacją przechodnią.

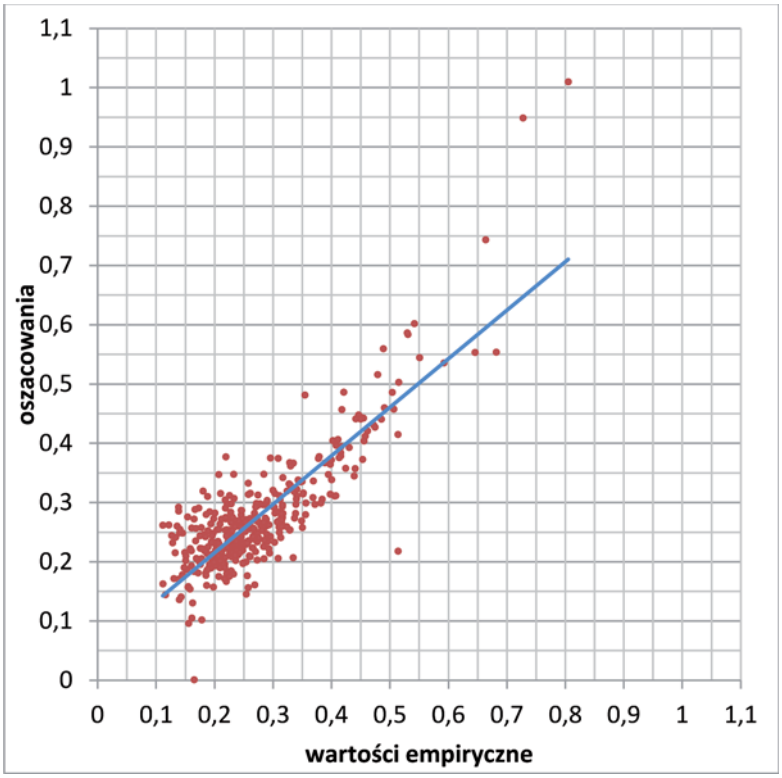
4.5 OSZACOWANIA ZMIENNEJ OBJAŚNIANEJ

Na podstawie otrzymanej w etapie piątym funkcji regresji dokonano oszacowań stosunku liczby osób pracujących do liczby osób w wieku produkcyjnym w powiatach według wzoru:

$$\hat{Y}_{ij} = 0,22040 + 0,50057 * Z_{1j} + \left(-2,48835 + 3,27839 * Z_{1j}\right) * X_{1ij} - 0,00688 * X_{2ij}.$$

Otrzymane oszacowania zbliżone są do wartości rzeczywistych dla większości powiatów. Duże wartości reszt otrzymano jedynie dla kilku jednostek, jednak poza nielicznymi obserwacjami odstającymi, wartości oszacowań badanej zmiennej charakteryzowały się dobrym dopasowaniem. świadczyć o tym może bliska liniowej zależność pomiędzy wartościami empirycznymi oraz oszacowaniami zmiennej przedstawiona na wykresie 3.

Dla połowy powiatów uzyskane z użyciem modelu dwupoziomowego oszacowanie różniło się od faktycznej wartości o nie więcej niż 11 procent. W przypadku jednej czwartej powiatów uzyskano błąd nie większy niż 5 procent, z kolei tylko dla 10 procent



Rysunek 3. Zależność pomiędzy empirycznymi wartościami zmiennej stosunek liczby osób pracujących do liczby osób w wieku produkcyjnym a ich oszacowania przy pomocy opracowanej funkcji regresji opartej na modelu dwupoziomowym, przekrój powiatów, Polska 2006

Źródło: Opracowanie własne

Tabela 10.

Parametry rozkładu błędów względnych oszacowań stosunku liczby osób pracujących do liczby osób w wieku produkcyjnym w powiatach przy zastosowaniu modelu dwupoziomowego

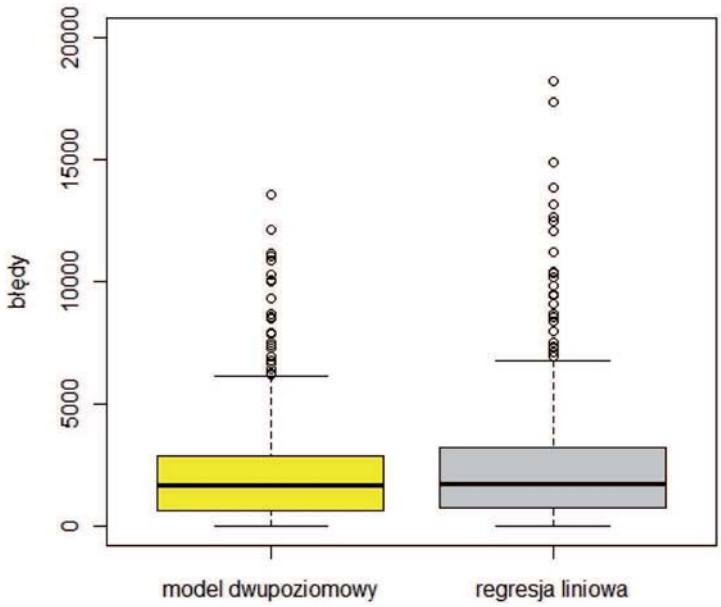
Procent	10	20	25	30	40	50	60	70	75	80	90
kwantyl	0,02	0,04	0,05	0,07	0,09	0,11	0,16	0,20	0,22	0,25	0,39

Źródło: Opracowanie własne

powiatów otrzymano oszacowanie różniące się od wartości empirycznej o więcej niż 39 procent (por. tab. 10).

W celu uzyskania oszacowań liczby osób pracujących w przekroju powiatów pomnożono oszacowany stosunek pracujących do ludności w wieku produkcyjnym przez liczbę ludności w wieku produkcyjnym. Tak otrzymane szacunki liczby osób pracujących porównano z wartościami teoretycznymi uzyskanymi w wyniku zastosowania klasycznej regresji liniowej. Przy zastosowaniu oszacowań opartych na opracowanym

modelu dwupoziomowym, w porównaniu ze zwykłą regresją liniową, uzyskano widoczną poprawę precyzji szacunku (por. rys. 4).



Rysunek 4. Rozkład błędów oszacowań liczby osób pracujących w powiatach przy zastosowaniu modelu dwupoziomowego oraz zwykłej regresji liniowej

Źródło: Opracowanie własne

Rozkłady błędów oszacowań otrzymanych przy zastosowaniu modelu dwupoziomowego oraz klasycznej regresji liniowej są do siebie zbliżone. Charakteryzują się silną asymetrią prawostronną. W przypadku obu metod charakterystyki rozkładu błędów, kwartyl pierwszy oraz drugi, są zbliżone, jednak kwartyl trzeci oraz największe nieodstające wartości są wyraźnie większe w przypadku klasycznej regresji liniowej. Wartość maksymalnego błędu uzyskanego przy zastosowaniu klasycznej regresji liniowej była o 29 procent większa od maksymalnego błędu otrzymanego przy zastosowaniu podejścia dwupoziomowego.

5. PODSUMOWANIE

Dzięki zastosowania metodologii modelowania dwupoziomowego udało się uzyskać znaczną poprawę jakości szacunku liczby osób pracujących w przekroju powiatów. świadczyć o tym może wzrost wartości funkcji wiarogodności modelu dwupoziomowego w stosunku do liniowej funkcji regresji z dwoma zmiennymi objaśniającymi (por. tab. 9).

Dzięki metodologii modelowania dwupoziomowego udało się uwzględnić zróżnicowanie poziomu badanej cechy pomiędzy województwami, co nie byłoby możliwe

przy zastosowaniu klasycznej funkcji regresji liniowej. Poprawiło to znacznie precyzję szacunku, ponieważ uwzględniono zróżnicowanie poziomu rozwoju gospodarczego w przekroju województw. Uzyskano także dodatkową wiedzę dzięki wykorzystaniu zmiennych objaśniających z drugiego poziomu, co także przyczyniło się do poprawy jakości szacunku. Potrzebę agregacji informacji dostępnych na różnych poziomach podkreślają w swojej pracy również Tadeusz Bołt, Kazimierz Krauze i Teodor Kulawczuk (por. Bołt, Krauze, Kulawczuk, 1985).

Warto podkreślić również, że skonstruowana zgodnie z omawianym schematem modelowania dwupoziomowego funkcja regresji zawsze charakteryzować musi się niemniejszą wiarogodnością od modelu regresji liniowej. Ponadto, jeżeli badane zmienne mają rzeczywiście strukturę dwupoziomową, tak jak w przedstawionym przykładzie, wiarogodność modelu dwupoziomowego jest znacznie większa. Jest tak dlatego, że model skonstruowany jest tak, że wprowadzenie dodatkowych informacji nie może pogorszyć jego wiarogodności, jeżeli komplikacja nie poprawia na zadanym poziomie istotności jakości szacunku nie jest wcale wprowadzana.

W związku z powyższym, w przypadku badania zmiennych o strukturze dwu lub wielopoziomowej, która charakteryzuje wiele ze zmiennych społeczno-ekonomicznych, stosowanie zwykłej regresji liniowej wiąże się z utratą części informacji, co z kolei oznacza gorszą precyzję szacunku.

W dalszych opracowaniach przedstawiona zostanie konstrukcja estymatora opartego na dwupoziomowej strukturze populacji oraz zmiennych w Statystyce Małych Obszarów w zakresie rynku pracy. Szczególna uwaga zostanie poświęcona sposobowi doboru zmiennych objaśniających z obu poziomów.

Uniwersytet Ekonomiczny w Poznaniu

LITERATURA

[1] Bliese P., (2012), *Multilevel Modeling in R (2.4) A Brief Introduction to R, the multilevel package and the nlme package*, Paul Bliese, April 10, http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf.
[2] Bołt T., Krauze K., Kulawczuk T., (1985), *Agregacja modeli ekonometrycznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.
[3] Goldstein H., (2003), *Multilevel Statistical Models*, 3rd edition, London: Edward Arnold.
[4] Harville D.A., (1974), *Bayesian Inference for Variance Components Using Only Error Contrasts*, Biometrika, 61, 383-385,
[5] Hox J., (2002), *Multilevel Analysis. Techniques and Applications*, Lawrence Erlbaum Associates, Publishers, London.
[6] Klimanek T., (2003), *Wielopoziomowa analiza struktury agrarnej gminy w systemie Geo-Info*, Praca doktorska napisana na Akademii Ekonomicznej w Poznaniu na Wydziale Zarządzania w Katedrze Statystyki, Poznań.
[7] Kopczewska K., Kopczewski T., Wójcik P., (2009), *Metody ilościowe w R Aplikacje ekonomiczne i finansowe*, CeDeWu.pl, Warszawa.
[8] Krzyśko M., (1996), *Statystyka matematyczna*, Wydawnictwo Naukowe UAM, Poznań.
[9] Lin X., (1997), *Variance Component Testing in Generalized Linear Models with Random Effects*, Biometrika, 84, 309-25.

[10] Pinheiro J.C., Bates D.M., (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag.
[11] Rao J.N.K., (2003), *Small Area Estimation*, Wiley & Sons, New York.
[12] Raudenbush S.W., Bryk A.S., (2002), *Hierarchical Linear Models. Applications and Data Analysis Methods*, Second Edition, Sage Publications, London_Thousand Oaks_New Delhi.
[13] Rydlewski J.P., (2009), *Estymatory największej wiarygodności w uogólnionych modelach regresji nieliniowej*, Praca doktorska napisana na Uniwersytecie Jagiellońskim na Wydziale Matematyki i Informatyki w Instytucie Matematyki, Kraków.
[14] Sakamoto Y., Ishiguro, M., and Kitagawa G., (1986), *Akaike Information Criterion Statistics*, D. Reidel Publishing Company.
[15] Schwarz G., (1978), *Estimating the Dimension of a Model*, Annals of Statistics, 6, 461-464.
[16] Twisk J.W.R., (2010), *Analiza wielopoziomowa – przykłady zastosowań Praktyczny podręcznik biostatyki i epidemiologii*, Oficyna Wydawnicza SGH, Warszawa.
[17] Węziak D., (2007), *Wielopoziomowe modelowanie regresyjne w analizie danych*, Wiadomości Statystyczne, 9 (556), 1-12.

MOŻLIWOŚCI ZASTOSOWANIA MODELOWANIA DWUPOZIOMOWEGO W BADANIACH EKONOMICZNYCH

S t r e s z c z e n i e

Głównym celem artykułu jest wykazanie przydatności metodologii modelowania dwupoziomowego w szacowaniu wartości zmiennych społeczno-ekonomicznych. W pierwszej części opracowania przedstawione zostaną etapy konstrukcji modelu uwzględniającego dwupoziomową strukturę badanej populacji oraz zmiennych. W części drugiej przedstawiono przykład zastosowania opisanej metodologii do szacowania liczby osób pracujących w przekroju powiatów. Jako jednostki drugiego poziomu przyjęto województwa.

Dzięki zastosowaniu metodologii modelowania dwupoziomowego możliwe jest uwzględnienie zróżnicowania poziomu badanej zmiennej oraz siły jej zależności ze zmiennymi objaśniającymi pomiędzy grupami. Ponadto, pozyskane zostały dodatkowe informacje dzięki wprowadzeniu zmiennych objaśniających z drugiego poziomu, czyli dotyczących całych grup.

W części drugiej jako zmienne objaśniające wykorzystane zostały miedzy innymi wyniki unikatowego badania przeprowadzonego w Urzędzie Statystycznym w Poznaniu, które dotyczyły przepływów związanych z zatrudnieniem. Głównym źródłem informacji tego badania są zasoby rejestrów podatkowych Ministerstwa Finansów. Dane te dotyczą roku 2006 i jest to pierwsza informacja od 1988 roku dotycząca dojazdów do pracy udostępniona przez GUS.

Przeprowadzone zostało porównanie jakości szacunków otrzymanych przy pomocy omawianego podejścia dwupoziomowego oraz klasycznej regresji liniowej. Otrzymane wyniki wskazują na przewagę modelu dwupoziomowego.

Słowa kluczowe: modelowanie dwupoziomowe, ANOVA, część losowa, dojazdy do pracy

THE POSSIBILITY OF TWO-LEVEL MODELING USED IN ECONOMIC STUDIES

A b s t r a c t

The main objective of this paper is to demonstrate the usefulness of two-level modeling methodology for estimating the socio-economic variables. In the first part of the paper one will present the model construction stages taking the two-level structure of population and variables into account. In the second part an example of using the above methodology for estimating the number of working people in cross-section of counties is presented. Province were chosen as second-level unit.

With the two-level modeling methodology one can include variation of the level of the considered variable and the strength of its dependencies with explanatory variables between the groups. Furthermore, additional information has been obtained by using the explanatory variables from the second level – concerning the entire group.

In the second part, as the explanatory variables, among others, the results of unique study in the Statistical Office in Poznań which concerned the flow of employees were used. The main source of information of this study are the fiscal records of the Ministry of Finance. These data concern the year 2006 and this has been the first information for commuting since 1988 provided by the Central Statistical Office.

The comparison of the quality of estimates obtained using two-level approach and the classical linear regression were conducted. The results show the advantage of two-level model.

Key words: two-level modeling, ANOVA, random component, commuting