Joanna Muszyńska

Modelowanie danych panelowych

Panel data models

Słowa kluczowe: dane panelowe, modele z dekompozycją, modele efektów stałych, modele efektów zmiennych

Key words: panel data, error component regression models, fixed effects models, random effects models

Synopsis

Celem artykułu było przedstawienie korzyści wynikających ze stosowania danych panelowych w modelowaniu ekonometrycznym. Omówiono także zasady budowy, estymacji i weryfikacji modeli panelowych. Rozważania teoretyczne poparto przykładem empirycznym.

1. Wprowadzenie

Podejmowanie decyzji ekonomicznych polega na właściwym wykorzystaniu dostępnych informacji. W procesie zarządzania przedsiębiorstwem coraz częściej stosuje się metody ekonometryczno-statystyczne, gdyż ułatwiają one podejmowanie decyzji strategicznych. Wspomaganie procesu decyzyjnego tymi metodami poprawia efektywność oraz skuteczność zarządzania.

Artykuł ma charakter metodologiczno-empiryczny. Jego celem, obok charakterystyki danych panelowych, jest przedstawienie korzyści wynikających z ich stosowania w modelowaniu ekonometrycznym. Omówione zostały także zasady budowy, estymacji i weryfikacji modeli panelowych. Rozważania teoretyczne poparto przykładem empirycznym,

opisującym zróżnicowanie wydatków konsumpcyjnych gospodarstw domowych w Polsce.

2. Istota danych panelowych

Modele ekonometryczne mogą być szacowane zarówno na podstawie szeregów czasowych jak i prób przekrojowych. Dane przekrojowe stanowią obserwacje zjawiska na różnych podmiotach, czynione w tym samym momencie czasu, przy czym kolejne obserwacje dotyczą różnych jednostek w pojedynczym okresie. Natomiast szereg czasowy przedstawia obserwacje zjawiska charakteryzującego jeden podmiot w różnych momentach w czasie.

Zależnie od rodzaju danych statystycznych wykorzystanych w procesie estymacji, modele ekonometryczne dzielą na statyczne i dynamiczne. W pierwszym przypadku, zmienna zależna jest zmienną losową, której realizacje stanowią dane przekrojowe, opisujące prawidłowości jej kształtowania się w tej samej jednostce czasu. Materiał statystyczny służący do budowy dynamicznych funkcji ma postać szeregów czasowych. W tym przypadku zmienna zależna jest procesem stochastycznym opisującym zmiany zjawiska w czasie.

W badaniach ekonometrycznych, „często w celu zwiększenia wielkości próby, łączy się dane w postaci szeregów czasowych z danymi przekrojowymi, tworząc tzw. próbę przekrojowo-czasową (próbę wymieszaną)”[1]. Takie połączenie obserwacji (*pooling*) jest możliwe pod warunkiem, że dysponujemy danymi dotyczącymi całej badanej populacji lub danymi na temat próby wyłonionej z tej populacji w sposób losowy. Próby wymieszane dostarczają informacji statystycznych o każdej jednostce badania w określonym przedziale czasu. Zgromadzone w ten sposób dane „są dostatecznie liczne, spełniają warunek względnej jednorodności

[1] Welfe W. (red.) 1977: Ekonometryczne modele rynku, tom I: Metody ekonometryczne. PWE Warszawa, str.82.

obserwacji statystycznych i umożliwiają dynamiczną analizę zjawisk ekonomicznych”[2].

Połączenie danych przekrojowych i szeregów czasowych przyczynia się przede wszystkim do zwiększenia liczby obserwacji. Dostarcza tym samym więcej informacji na temat badanych zjawisk, a to ułatwia ustalenie istniejących między nimi zależności oraz ich ocenę. Korzystanie z prób wymieszanych sprzyja także dokonywaniu bardziej wnikliwych, szczegółowych analiz, których nie można przeprowadzić na innych rodzajach danych statystycznych. Łączenie obserwacji przekrojowych i czasowych „pozwala na identyfikację i pomiar efektów, nie dających się zaobserwować na typowych danych przekrojowych lub typowych szeregach czasowych”[3]. *Pooling* umożliwia jednoczesne uwzględnienie w modelu informacji o dynamice czasowej jak i indywidualnej specyfice badanych obiektów.

Zastosowanie prób wymieszanych do budowy modeli ekonometrycznych dostarcza również wielu korzyści pod względem estymacji parametrów strukturalnych. „W takiej sytuacji, gdy dane są generowane przez bardzo podobne procesy ekonomiczne, które mogą być opisane przy pomocy tego samego modelu, można połączyć dane dotyczące wszystkich badanych obiektów. Estymacja jest wtedy bardziej efektywna niż estymacja każdej jednostki osobno”[4]. Wzrost wielkości próby zwiększa dyspersję zmiennych, a tym samym zmniejsza ich współliniowość. Większa liczba obserwacji pozwala na estymację większej liczby parametrów strukturalnych modelu, przyczynia się także do zwiększenia dokładności

[2] Grabiński T., Malina A., Zeliaś A. 1990: Metody analizy danych empirycznych na podstawie szeregów przekrojowo-czasowych. AE Kraków, str.79.
[3] Baltagi B.H. 2001: Econometric Analysis of Panel Data. John Wiley & Sons Ltd., Chichester, str.7.
[4] Dańska B. 1995: Wybrane metody estymacji modeli opartych na danych panelowych. UŁ Łódź, str.3.

otrzymanych szacunków. Wzrost liczby stopni swobody modelu ułatwia jego weryfikację statystyczną.

Szczególny przypadek prób przekrojowo-czasowych stanowią dane panelowe. Są to obserwacje statystyczne dotyczące ustalonej grupy badanych jednostek, w kolejnych momentach czasu. Tworzą więc szeregi czasowe danych dla stałej próby przekrojowej. W ten sposób dostarczają one „sekwencyjnych obserwacji dla konkretnych jednostek dla wielu okresów. Pozwala to na rozróżnienie efektów indywidualnych od efektów powodowanych przez czynniki zewnętrzne – możliwe staje się kontrolowanie wpływu indywidualnego, wewnętrznego zróżnicowania jednostek”[5]. Różnice między panelem a próbą przekrojowo-czasową są na tyle subtelne, że często nazwy te stosowane są zamiennie. „Dane panelowe należy odróżnić od prób przekrojowo-czasowych, dla których dysponujemy szeregiem różnych prób przekrojowych dla kolejnych momentów czasu”[6]. Choć panele, tak jak próby wymieszane, mają charakter zarówno próby przekrojowej jak i szeregu czasowego, to służą głównie analizom przekrojowym. „(...) próby te są szerokie, ale charakterystycznie krótkie. Heterogeniczność jednostek jest tu integralną częścią analizy, a często stanowi nawet jej główne zagadnienie”[7]. Zbyt długi okres badania może prowadzić do „starzenia się” danych i utraty reprezentatywności próby, dlatego oryginalne panele (*genuine panels*) charakteryzują się tym, że „liczba obserwowanych obiektów jest bardzo duża w stosunku do liczby punktów w czasie”[8].

[5] Ciecieląg J., Tomaszewski A. 2003: Ekonometryczna analiza danych panelowych. WNE UW Warszawa, str. 5.
[6] http://www.jmyc.republica.pl
[7] Greene W. H. 2000: Econometric Analysis. Prentice-Hall, Inc., New Jersey, str.557.
[8] Suchecki B. (red.) 2000: Dane panelowe i modele wielowymiarowe w badaniach ekonomicznych, tom I: Dańska B. 2000: Przestrzenno-czasowe modelowanie zmian w działalności produkcyjnej w Polsce. Zastosowanie modeli panelowych. Absolwent Łódź, str.7.

Stosowanie prób panelowych do budowy modeli ekonometrycznych pozwala na uwzględnienie zróżnicowania badanych jednostek i obserwacji przemian w czasie pojedynczych obiektów, przy jednoczesnej agregacji danych. „Estymacja modeli na danych zagregowanych często powoduje rozmycie się teoretycznych własności zależności ekonomicznych. (...) Użycie danych panelowych umożliwia więc estymację parametrów modeli behawioralnych, które łatwiej interpretować w ramach teorii ekonomii”[9].

3. Budowa modeli dla danych panelowych

Modele ekonometryczne szacowane na podstawie danych panelowych tworzą grupę tzw. modeli z dekompozycją (error component regression models), nazywanych krótko modelami panelowymi. Modele wielorównaniowe estymowane w oparciu o dane panelowe są jeszcze stosunkowo rzadko wykorzystywane w praktyce. W badaniach empirycznych najczęściej stosuje się dwa rodzaje modeli jednorównaniowych: modele z dekompozycją przestrzenną (one-way error component regression model), określane także jako modele błędu jednokierunkowego lub modele jednoczynnikowe oraz modele z dekompozycją przestrzenną i czasową (two-way error component regression model), nazywane modelami błędu dwukierunkowego albo modelami dwuczynnikowymi.

Budowa modeli błędu jednokierunkowego oparta jest na założeniu, że jednostki badania (grupy jednostek) różnią się od siebie pewnymi charakterystykami, które dla konkretnego podmiotu pozostają stałe w czasie. Liniowy model jednoczynnikowy można więc zapisać następująco:

$$y_{it} = \alpha + X'_{it}\beta + u_{it}, \qquad dla \quad i = 1,....,N; \quad t = 1,...,T; \quad (3.1)$$

$$u_{it} = \mu_i + \nu_{it}, \qquad (3.2)$$

[9] http://www.jmyc.republica.pl

gdzie:

i, t – indeksy oznaczające odpowiednio jednostkę badania i czas,

X_{it} – wektor obserwacji na zmiennych egzogenicznych,

α, β - parametry strukturalne modelu,

u_{it} – składnik losowy modelu,

μ_i – nieobserwowalny i nieuwzględniony w równaniu regresji efekt indywidualny[10] właściwy dla danej jednostki badania,

ν_{it} – pozostała, czysto losowa część składnika losowego.

Efekty grupowe μ_i interpretowane są jako indywidualne charakterystyki jednostek badania, które nie podlegają zmianom w czasie.

Zależnie od sposobu traktowania efektów indywidualnych, jako zmienne losowe lub jako wielkości nielosowe, w grupie modeli błędu jednokierunkowego można wyróżnić modele efektów stałych i modele efektów zmiennych. Pierwsze z nich bazują na założeniu, że efekty grupowe mają charakter stałych parametrów. Ze względu na powyższe założenie modele te nazywane są także modelami ze sztucznymi zmiennymi lub modelami z dekompozycją wyrazu wolnego. Modele efektów zmiennych określa się w literaturze terminem modeli ze składnikami błędu lub modeli z dekompozycją składnika losowego. Opierają się one na założeniu, że efekty grupowe są zmienną losową o znanym rozkładzie.

Modele błędu dwukierunkowego uwzględniają istnienie nie tylko efektów grupowych, ale i efektów czasowych. Zakładają więc możliwość, że w danym momencie to samo zaburzenie losowe dotyka wszystkie jednostki badania. Budowa liniowych modeli dwuczynnikowych jest analogiczna do budowy modeli jednoczynnikowych (3.1), a jedynie składnik losowy ulega dekompozycji na trzy elementy składowe:

[10] Ponieważ efekt indywidualny może wynikać z przynależności jednostki do *i*-tej grupy często bywa również określany terminem efekt grupowy.

$$u_{it} = \mu_i + \lambda_t + \nu_{it}, \qquad (3.3)$$

gdzie:

λ_t - nieobserwowalny i nieuwzględniony w równaniu regresji efekt specyficzny dla czasu.

Efekty grupowe μ_i nie podlegają zmianom w czasie, natomiast efekty czasowe λ_t pozostają stałe dla wszystkich jednostek badania.

Podobnie, jak wśród modeli jednoczynnikowych, w grupie modeli z błędem dwukierunkowym, w zależności od założeń dotyczących efektów grupowych i czasowych, wyróżnia się modele efektów stałych i modele efektów zmiennych.

Wybór między modelami jedno- i dwuczynnikowymi jest przede wszystkim uwarunkowany celem prowadzonego badania. Natomiast założenie o stałości efektów grupowych i czasowych lub o ich losowości pociąga za sobą konsekwencje w doborze metody estymacji parametrów modelu, a także dalszych możliwości wnioskowania. Modele z czynnikami stałymi stosuje się najczęściej dla tzw. paneli długich, o małej liczbie jednostek i długim okresie badania, gdy możemy być pewni, że różnice pomiędzy jednostkami da się uchwycić jako różnice w wyrazie wolnym. „Modele efektów stałych stanowią odpowiednią specyfikację, jeżeli proces badawczy koncentruje się na wybranej grupie podmiotów, a wynikające z badania wnioski ograniczają się jedynie do badanych jednostek”[11]. Bazują one na założeniu, że różnice pomiędzy jednostkami badania mogą być przedstawione poprzez różne wartości stałej w modelu. Oznacza to, że dla każdej z nich ocena wyrazu wolnego będzie osiągała inną wartość. W ten sposób uwzględniony zostanie wpływ wszystkich niezmiennych w czasie czynników, specyficznych dla danej jednostki badania. „Model taki może

[11] Baltagi B.H. 2001: Econometric … op.cit., str.12.

być dalej stosowany tylko dla jednostek uczestniczących w badaniu, a nie dla dodatkowych jednostek spoza próby”[12].

Jeśli zaś elementy w próbie zostały wylosowane z dużej populacji, należy założyć, że efekt grupowy jest realizacją pewnej zmiennej losowej o znanym rozkładzie. „Modele efektów zmiennych stanowią odpowiednią specyfikację, jeżeli jednostki badania zostały wybrane z całej populacji w sposób losowy, (...) a panel zaprojektowano jako próbę reprezentatywną, na podstawie której wyciągane są wnioski dotyczące całej populacji”[13].

Zdaniem niektórych ekonometryków „modele z dekompozycją składnika losowego są bardziej użyteczne, niż modele z dekompozycją wyrazu wolnego, gdyż do szacowania parametrów tych pierwszych oprócz różnic między grupami wykorzystuje się różnice między okresami – źródło, które jest całkowicie eliminowane w modelach z dekompozycją wyrazu wolnego. Taki pogląd uzasadniony jest dodatkowo tym, że efekty grupowe, tak samo jak składnik losowy, są „miarą niewiedzy” konstruktora modelu, nie ma zatem powodu, by jedno źródło niewiedzy traktować jako losowe, a inne jako nielosowe”[14]. Dodatkowo, wprowadzenie do modelu szeregu sztucznych zmiennych powoduje znaczącą utratę stopni swobody. „Z drugiej strony, modele efektów stałych mają jedną poważną zaletę. Nie istnieje żadne usprawiedliwienie dla uznania, że efekty indywidualne nie są skorelowane z pozostałymi regresorami, jak zakładają to modele efektów zmiennych. Dlatego podejście losowe może być przyczyną utraty zgodności estymatora z powodu pominiętych zmiennych”[15].

[12] Greene W. H. 2000: Econometric … op.cit., str.567.
[13] Baltagi B.H. 2001: Econometric … op.cit., str.15.
[14] Suchecki B. (red.) 2000: Dane panelowe ...op.cit., str.19.
[15] Greene W. H. 2000: Econometric… op.cit., str.576.

4. Estymacja parametrów modeli panelowych

Szacowanie parametrów modeli na podstawie danych panelowych wymaga stosowania metod, które umożliwią wyodrębnienie różnic pomiędzy obiektami w tym samym okresie, jak i pomiędzy różnymi okresami dla tego samego obiektu. Wybór procedury estymacyjnej jest uwarunkowany założeniami dotyczącymi stałości lub losowości efektów grupowych i czasowych. Nie zależy on natomiast od stopnia dekompozycji.

4.1 Modele efektów stałych

Model jednoczynnikowy, przedstawiony w postaci równaniowej za pomocą wzorów (3.1) i (3.2), można zapisać łącznie w postaci macierzowej jako:

$$y = \alpha i_{NT} + X\beta + u = Z\delta + u, \tag{4.1}$$

$$u = Z_{\mu}\mu + \nu, \tag{4.2}$$

gdzie:

y – wektor zmiennych endogenicznych o NT współrzędnych,

X – macierz zmiennych egzogenicznych o wymiarach (NTxK),

i_{NT} – wektor jedynek o NT wymiarach,

Z_{μ} - macierz selekcjonująca, określana jako macierz sztucznych zmiennych.

Zakładając, że μ jest wektorem nielosowych parametrów, ν wektorem składników losowych o jednakowych, niezależnych rozkładach: $\nu_{it}:N(0,\sigma_{\nu}^2)$, a X_{it} są niezależne od ν_{it} dla wszystkich i, t otrzymujemy model błędu jednokierunkowego ze sztucznymi zmiennymi:

$$y = \alpha i_{NT} + X\beta + Z_{\mu}\mu + \nu = Z\delta + Z_{\mu}\mu + \nu. \tag{4.3}$$

Parametry δ i μ powyższego równania można oszacować za pomocą KMNK. By w procesie estymacji uniknąć problemu ścisłej współliniowości

zmiennych, należy usunąć z modelu jedną ze zmiennych zero-jedynkowych lub stałą, bądź też narzucić na wektor parametrów μ ograniczenie, np. $\sum_{i=1}^{N} \mu_i = 0$. Uzyskany w ten sposób estymator nazywamy estymatorem metody najmniejszych kwadratów ze sztucznymi zmiennymi (LSDV – Least Squares Dummy Variables).

Analogiczna procedura estymacji dotyczy modeli błędu dwukierunkowego. Model dwuczynnikowych ma postać:

$$y = \alpha i_{NT} + X\beta + Z_{\mu}\mu + Z_{\lambda}\lambda + \nu, \tag{4.4}$$

gdzie:

$$u = Z_{\mu}\mu + Z_{\lambda}\lambda + \nu \tag{4.5}$$

jest składnikiem losowym modelu, a macierz Z_{λ} dana jest równaniem:

$$Z_{\lambda} = i_N \otimes I_T \tag{4.6}$$

Zakładamy, że wektory μ i λ są wektorami nielosowych parametrów, ν wektorem składników losowych o jednakowych, niezależnych rozkładach: ν_{it}:$N(0, \sigma_{\nu}^2)$, a X_{it} są niezależne od ν_{it} dla wszystkich *i, t*. Tak zdefiniowany model staje się dwuczynnikowym modelem z dekompozycją wyrazu wolnego o N+T-2 sztucznych zmiennych. Parametry powyższego modelu szacujemy przy pomocy estymatora LSDV. Aby uniknąć problemu ścisłej współliniowości usuwamy stałą lub po jednej ze zmiennych zero-jedynkowych, z obu grup tych zmiennych, bądź też nakładamy na parametry przy nich odpowiednie ograniczenia.

4.2 Modele efektów zmiennych

Modele efektów zmiennych mają zastosowanie, gdy próba, na podstawie której został zbudowany panel, jest próbą reprezentatywną. Model jednoczynnikowy, przedstawiony za pomocą równań (4.1) i (4.2), oparty na

założeniach, że μ i ν są niezależne, μ_i:$N(0,\sigma_\mu^2)$, ν_{it}:$N(0,\sigma_\nu^2)$, a X_{it} są niezależne od μ_i i ν_{it} dla wszystkich *i*, *t* staje się modelem z dekompozycją składnika losowego. Macierz wariancji – kowariancji składnika losowego ma postać:

$$\begin{aligned}\Omega &= E(uu') = Z_\mu E(\mu\mu')Z'_\mu + E(\nu\nu') = \\ &= \sigma_\mu^2(I_N \otimes J_T) + \sigma_\nu^2(I_N \otimes I_T).\end{aligned} \tag{4.7}$$

Oznacza to stałość wariancji składnika losowego modelu dla wszystkich *i* oraz *t:*

$$\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\nu^2. \tag{4.8}$$

Zatem błąd losowy modelu jest homoskedatyczny, ale macierz Ω jest niediagonalna:

$$\text{cov}(u_{it}, u_{js}) = \begin{cases} \sigma_\mu^2 + \sigma_\nu^2 & dla \quad i = j, t = s, \\ \sigma_\mu^2 & dla \quad i = j, t \neq s, \\ 0 & dla \quad pozostalych. \end{cases} \tag{4.9}$$

W modelu istnieje korelacja w czasie między składnikami losowymi dotyczącymi tych samych obiektów, nie występuje natomiast korelacja składników losowych różnych obiektów w różnych okresach czasu. Do estymacji parametrów powyższego modelu należy więc zastosować UMNK.

Dwuczynnikowe modele efektów zmiennych oparte są na założeniach, że μ, λ i ν są niezależne, μ_i:$N(0,\sigma_\mu^2)$, λ_t:$N(0,\sigma_\lambda^2)$, ν_{it}:$N(0,\sigma_\nu^2)$, a X_{it} są niezależne od μ_i, λ_t i ν_{it} dla wszystkich *i*, *t*. Macierz wariancji – kowariancji składnika losowego ma w tym przypadku postać:

$$\begin{aligned}\Omega &= E(uu') = Z_\mu E(\mu\mu')Z'_\mu + Z_\lambda E(\lambda\lambda')Z'_\lambda + \sigma_\nu^2 I_{NT} = \\ &= \sigma_\mu^2(I_N \otimes J_T) + \sigma_\lambda^2(J_N \otimes I_T) + \sigma_\nu^2(I_N \otimes I_T).\end{aligned} \tag{4.10}$$

Składnik losowy jest zatem homoskedastyczny o wariancji*:*

$$\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\lambda^2 + \sigma_\nu^2 \tag{4.11}$$

dla wszystkich *i* oraz *t,* zaś kowariancję między błędami losowymi można przedstawić*:*

$$\text{cov}(u_{it}, u_{js}) = \begin{cases} \sigma_\mu^2 & dla \quad i = j, t \neq s, \\ \sigma_\lambda^2 & dla \quad i \neq j, t = s, \\ 0 & dla \quad pozostalych. \end{cases} \tag{4.12}$$

Ponieważ macierz Ω jest niediagonalna, do szacowania parametrów modelu należy zastosować UMNK.

5. Testy statystyczne dla modeli panelowych

Modele panelowe poza wyborem odpowiedniej procedury estymacyjnej wymagają również przeprowadzenia testów, których celem jest weryfikacja przyjętych założeń o stałości lub losowości efektów grupowych i czasowych.

5.1 Test F (Chowa)

Celem stosowania testu F jest sprawdzenie łącznej istotności sztucznych zmiennych. Dla modeli jednoczynnikowych hipoteza zerowa ma postać:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0. \tag{5.1}$$

Statystyką służącą do jej weryfikacji jest empiryczna statystyka F obliczona według wzoru:

$$F_1 = \frac{u'u - u'_W u_W}{u'_W u_W} \cdot \frac{NT - N - K}{N - 1}, \tag{5.2}$$

gdzie:

u - wektor reszt modelu bez sztucznych zmiennych (tzw. *pooled regression*),

u_W – wektor reszt uzyskanych przy zastosowaniu estymatora LSDV.

Powyższa statystyka ma rozkład F o (N-1) i (NT-N-K) stopniach swobody. Jeśli wartość F_0 jest większa od odpowiedniej wartości krytycznej odrzuca się H_0 na korzyść hipotezy alternatywnej.

W modelach dwuczynnikowych występują dwie grupy zmiennych sztucznych: zmienne dotyczące efektów grupowych i zmienne opisujące efekty czasowe. Weryfikacji może więc podlegać łączna istotność wszystkich zmiennych zero-jedynkowych lub istotność zmiennych sztucznych należących do jednej z grup. Hipoteza zerowa mówiąca o łącznej nieistotności wszystkich sztucznych zmiennych ma postać:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0 \quad \wedge \quad \lambda_1 = \ldots = \lambda_{T-1} = 0 . \qquad (5.3)$$

Do weryfikacji powyższej hipotezy stosowana jest statystyka F:

$$F_2 = \frac{u'u - u'_W u_W}{u'_W u_W} \cdot \frac{(N-1)(T-1) - K}{N + T - 2} . \qquad (5.4)$$

Statystyka ma rozkład F o (N+T-2) i [(N-1)(T-1)-K] stopniach swobody. Jeśli wartość F_2 przewyższa wartość krytyczną, hipoteza zerowa jest odrzucana.

W przypadku badania istotności efektów grupowych hipoteza zerowa ma postać:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0 \quad \wedge \quad \lambda_t \neq 0, \quad dla \quad t = 1, \ldots = T - 1. \qquad (5.5)$$

Podobnie można również zweryfikować hipotezę o istotności efektów czasowych. Hipoteza zerowa ma wtedy postać:

$$H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_{T-1} = 0 \quad \wedge \quad \mu_i \neq 0, \quad dla \quad i = 1, \ldots = N - 1. \qquad (5.6)$$

Empiryczna statystyka F, służąca do weryfikacji powyższych hipotez dana jest wzorem:

$$F_3 = \frac{u_T' u_T - u_W' u_W}{u_W' u_W} \cdot \frac{(N-1)(T-1) - K}{N-1}, \tag{5.7}$$

gdzie:

u_T – wektor reszt modelu uwzględniającego zmienne zero-jedynkowe opisujące efekty czasowe / grupowe (zależnie od postawionej hipotezy).

Wartość statystyki F_3 porównuje się z wartością krytyczną odczytaną z rozkładu F dla (N-1) i [(N-1)(T-1)-K] stopni swobody. Jeśli jest ona większa następuje odrzucenie hipotezy zerowej na korzyść alternatywnej.

5.2 Test Breuscha-Pagana dla modeli z dekompozycją składnika losowego

Test Breuscha-Pagana, oparty na mnożniku Lagrange'a, pozwala na weryfikację hipotezy, że model z dekompozycją składnika losowego jest statystycznie lepszy od modelu, w którym nie wyróżniono efektów grupowych i/lub czasowych.

W przypadku weryfikacji hipotezy o łącznej nieistotności efektów grupowych i efektów czasowych hipoteza zerowa ma postać:

$$H_0 : \sigma_\mu^2 = \sigma_\lambda^2 = 0. \tag{5.9}$$

Do weryfikacji powyższej hipotezy wykorzystuje się mnożnik Lagrange'a wyznaczony zgodnie ze wzorem:

$$LM = LM_1 + LM_2 , \tag{5.10}$$

gdzie:

$$LM_1 = \frac{NT}{2(T-1)} \left[1 - \frac{u'(I_N \otimes J_T)u}{u'u} \right]^2, \tag{5.11}$$

$$LM_2 = \frac{NT}{2(N-1)} \left[1 - \frac{u'(J_N \otimes I_T)u}{u'u} \right]^2, \tag{5.12}$$

u – reszty modelu, w którym nie wyróżniono efektów grupowych i czasowych.

Statystyka *LM* ma rozkład χ^2 z dwoma stopniami swobody. Jeśli wartość statystyki *LM* przewyższa wartość krytyczną efekty grupowe i czasowe są statystycznie istotne.

Test Breuscha-Pagana można zastosować również to oddzielnego testowania istotności efektów grupowych i efektów czasowych. W pierwszym przypadku hipoteza zerowa będzie miała postać:

$$H_0 : \sigma_\mu^2 = 0, \tag{5.13}$$

a jej weryfikacja opierała się będzie wyłącznie o wartość statystyki LM_1.

Natomiast hipoteza zerowa:

$$H_0 : \sigma_\lambda^2 = 0, \tag{5.14}$$

mówiąca o nieistotności efektów czasowych, weryfikowana jest wyłącznie na podstawie statystyki LM_2. Każda z powyższych statystyk ma rozkład χ^2 z jednym stopniem swobody. W obu przypadkach H_0 jest odrzucana jeśli wartości LM_1 (LM_2) jest większa od wartości krytycznej.

5.3 Test Hausmana

Stosowanie modeli z efektami stałymi wymaga wprowadzenia do równania regresji dodatkowych, sztucznych zmiennych, co powoduje utratę dużej liczby stopni swobody i może prowadzić do pogorszenia własności estymatorów parametrów α i μ_i. Natomiast modele z efektami zmiennymi opierają się na założeniu, że efekty grupowe μ_i nie są skorelowane ze zmiennymi egzogenicznymi modelu. Warunek ten nie zawsze jest spełniony, a pominięcie istotnych zmiennych może powodować heteroskedatyczność składnika losowego modelu.

Testem pozwalającym na sprawdzenie czy efekty grupowe są ortogonalne w stosunku do macierzy *X* jest test Hausmana. Bazuje on na własnościach estymatorów UMNK i wewnątrzobiektowego. Jeśli w modelu występuje korelacja między składnikiem losowym a zmiennymi egzogenicznymi, to estymator $\hat{\beta}_{UMNK}$ staje się obciążony i niezgodny. Tymczasem przekształcenie stosowane przy konstrukcji estymatora $\tilde{\beta}_W$ usuwa efekty grupowe stanowiące przyczynę korelacji, a estymator pozostaje zgodny i nieobciążony.

Hipoteza zerowa, mówiąca o braku korelacji między składnikiem losowym modelu a zmiennymi egzogenicznymi ma postać:

$$H_0 : E(u_{it} / X_{it}) = 0, \tag{5.15}$$

i gdy jest ona prawdziwa oba estymatory $\hat{\beta}_{UMNK}$ i $\tilde{\beta}_W$ są zgodne. Jeśli natomiast $E(u_{it} / X_{it}) \neq 0$, to estymator wewnątrzobiektowy nadal pozostaje zgodny, podczas gdy estymator UMNK traci zgodność. Statystyka weryfikacyjna testu opiera się na różnicy występującej między estymatorami, jeśli H_0 jest nieprawdziwa.

$$H = \hat{q}_1' [\text{var}(\hat{q}_1)]^{-1} \hat{q}_1 \tag{5.16}$$

gdzie:

$$\hat{q}_1 = \hat{\beta}_{UMNK} - \tilde{\beta}_W, \tag{5.17}$$

$$\text{var}(\hat{q}_1) = \text{var}(\tilde{\beta}_W) - \text{var}(\hat{\beta}_{UMNK}). \tag{5.18}$$

Statystyka *H* ma rozkład χ^2 z *k* stopniami swobody, gdzie *k* jest liczbą zmiennych w macierzy X.

6. Zróżnicowanie wydatków konsumpcyjnych gospodarstw domowych w Polsce

W celu zobrazowania sposobu wykorzystania modeli panelowych w praktyce, postawiono hipotezę o przestrzennym i czasowym zróżnicowaniu wydatków konsumpcyjnych gospodarstw domowych w Polsce. Jej weryfikacji dokonano bazując na informacjach statystycznych o budżetach gospodarstw domowych gromadzonych i agregowanych przez Główny Urząd Statystyczny. Badanie przeprowadzono w oparciu o wielkości realne. W charakterze deflatorów wykorzystano wskaźniki cen dóbr konsumpcyjnych.

W roli narzędzia badawczego wykorzystano jedno- i dwuczynnikowe modele panelowe ze sztucznymi zmiennymi oraz z dekompozycją składnika losowego. Zmienną zależną zdefiniowano jako realną wartość wydatków na towary i usługi konsumpcyjne, przypadającą na 1 osobę w gospodarstwie domowym. W roli zmiennej egzogenicznej wykorzystano przeciętne miesięczne dochody na osobę w gospodarstwie domowym W modelach jednoczynnikowych podstawą dekompozycji wyrazu wolnego / składnika losowego była lokalizacja gospodarstwa domowego, natomiast w modelach dwuczynnikowych lokalizacja oraz czas.

Każdy z modeli estymowany był pięciokrotnie: jako model bez dekompozycji (pooled regression), jako model jednoczynnikowy ze sztucznymi zmiennymi i z dekompozycją składnika losowego oraz jako model dwuczynnikowy z efektami stałymi i z efektami losowym. Ze względu na przejrzystą interpretację wyników, równania szacowano w

dwóch postaciach analitycznych: liniowej i potęgowej[16]. Szacowane równania modeli miały postać:

$$wydatek = \alpha + \beta_1 dochód + \beta_2 czas + \mu_i + \nu_{it}, \quad (6.1)$$

$$\ln(wydatek) = \ln\alpha + \beta_1 \ln(dochód) + \beta_2 czas + \mu_i + \nu_{it}, \quad (6.2)$$

$$wydatek = \alpha + \beta_1 dochód + \mu_i + \lambda_t + \nu_{it}, \quad (6.3)$$

$$\ln(wydatek) = \ln\alpha + \beta_1 \ln(dochód) + \mu_i + \lambda_t + \nu_{it}, \quad (6.4)$$

Występujące w modelach z dekompozycją wyrazu wolnego zmienne zero-jedynkowe, określające miejsce zamieszkania członków gospodarstwa domowego, oznaczono następująco: *du_1*– województwo dolnośląskie, *du_2*– kujawsko-pomorskie, *du_3*– lubelskie, *du_4*– lubuskie, *du_5*– łódzkie, *du_6*– małopolskie, *du_7*– mazowieckie, *du_8*– opolskie, *du_9*– podkarpackie, *du_10*– podlaskie, *du_11*– pomorskie, *du_12*– śląskie, *du_13*– świętokrzyskie, *du_14*– warmińsko-mazurskie, *du_15*– wielkopolskie i *du_16*– zachodniopomorskie.

W związku z tym, że wszystkie estymowane modele charakteryzowały się wysokim stopniem zgodności danych empirycznych i teoretycznych, w artykule zaprezentowano wyniki empiryczne modeli jednoczynnikowych w postaci liniowej oraz modeli dwuczynnikowych w postaci potęgowej. Oceny parametrów zamieszczono w tablicach 6.1 i 6.2.

W modelu z efektami stałymi, oceny parametrów przy zmiennych ‘dochód’ oraz ‘czas’ wyniosły odpowiednio 0,63 i –2,20. Oznacza to, że wraz ze wzrostem dochodu przypadającego na 1 osobę w gospodarstwie domowym o 1 PLN, wartość wydatków konsumpcyjnych wzrastała średnio o 0,63 PLN, natomiast co roku obserwowano spadek wydatków przeciętnie 0,20 PLN

[16] W modelach potęgowych czas, efekty grupowe oraz składnik losowy nakładano wykładniczo.

Tablica 6.1. *Oceny parametrów jednoczynnikowego modelu wydatków konsumpcyjnych*

Zmienna	model ze sztucznymi zmiennymi		model z dekompozycją składnika losowego	
	wartość oceny	statystyka t	wartość oceny	statystyka t
wyraz wolny	-	-	92,47	4,20
du_1	169,84	4,79	-	-
du_2	141,24	4,42	-	-
du_3	144,63	4,83	-	-
du_4	176,13	5,26	-	-
du_5	170,70	4,85	-	-
du_6	160,36	4,77	-	-
du_7	177,84	4,13	-	-
du_8	170,19	5,05	-	-
du_9	147,50	5,24	-	-
du_10	156,19	5,15	-	-
du_11	154,03	4,21	-	-
du_12	172,99	4,71	-	-
du_13	147,35	5,05	-	-
du_14	141,05	4,63	-	
du_15	144,93	4,27	-	-
du_16	171,16	4,88	-	-
dochód	0,63	8,99	0,77	17,40
czas	-2,20	-2,15	-2,76	-3,87
R^2	0,914	-	0,736	-
var (μ_i)	-	-	105,67	-
var (v_{it})	251,32	-	211,21	-

Źródło: Obliczenia własne

Dla modelu ze sztucznymi zmiennymi oszacowano 16 wyrazów wolnych specyficznych dla każdego województwa. Test Chowa potwierdził łączną istotność zmiennych zero-jedynkowych. Oznacza to, że model z dekompozycją wyrazu wolnego lepiej opisał wydatki konsumpcyjne gospodarstw domowych, niż model bez dekompozycji, oparty na tych samych danych. Wartości ocen wyrazu wolnego, dla poszczególnych województw, wahały się od 141,05 PLN – dla gospodarstw zamieszkałych w województwie warmińsko-mazurskim – do 177,84 PLN w Mazowieckiem.

Oceny parametrów, w modelu z efektami losowymi, informują, że wzrostowi dochodu o 1 PLN odpowiada wzrost wydatków konsumpcyjnych przeciętnie o 0,77 PLN. Rokrocznie następował jednak spadek wydatków średnio o 2,76 PLN.

W modelach z dekompozycją składnika losowego efekty grupowe traktowane są jako zmienne losowe. Ich zróżnicowanie odzwierciedla odpowiednia składowa wariancji składnika losowego var(μ_i). Testem weryfikującym hipotezę o statystycznej istotności efektów grupowych w modelach z dekompozycją składnika losowego jest test mnożników Lagarnge'a Breuscha-Pagana. Wartość statystyki dla modelu wydatków wyniosła LM=12,239 i była większa od wartości krytycznej z rozkładu χ^2 na poziomie istotności α=0,05. Oznacza to, że model z efektami losowymi także lepiej wyjaśnia kształtowanie się badanego zjawiska niż model pooled regression.

Do porównania modeli ze sztucznymi zmiennymi i z dekompozycją składnika losowego wykorzystano test Hausmana. Obliczona wartość statystyki *H* wyniosła *H*=13,10, co oznacza, że należy odrzucić hipotezę zerową, mówiącą o braku korelacji między składnikiem losowym a zmiennymi egzogenicznymi modelu. Test Hausmana potwierdza stałość efektów grupowych. Oznacza to, że zróżnicowanie wydatków konsumpcyjnych gospodarstw domowych wynika z miejsca zamieszkania jego członków.

Ocena parametru przy zmiennej 'dochód', w modelu ze sztucznymi zmiennymi, wyniosła 0,54. Jednoprocentowemu wzrostowi dochodu przypadającego na 1 osobę w gospodarstwie domowym, odpowiadał zatem wzrost wydatków na ten cel przeciętnie o 0,54%.

Tablica 6.2. *Oceny parametrów dwuczynnikowego modelu wydatków konsumpcyjnych*

Zmienna	model ze sztucznymi zmiennymi		model z dekompozycją składnika losowego	
	wartość oceny	statystyka t	wartość oceny	statystyka t
wyraz wolny	-	-	1,249	4,23
du_1	-	-	-	-
du_2	-0,074	-8,55	-	-
du_3	-0,075	-5,30	-	-
du_4	0,006	1,31	-	-
du_5	0,002	3,48	-	-
du_6	-0,025	-5,70	-	-
du_7	0,043	2,66	-	-
du_8	-0,005	-1,08	-	-
du_9	-0,076	-3,98	-	-
du_10	-0,046	-3,60	-	-
du_11	-0,027	-10,3	-	-
du_12	0,011	3,79	-	-
du_13	-0,072	-4,45	-	-
du_14	-0,081	-6,44	-	
du_15	-0,057	-15,7	-	-
du_16	0,002	2,85	-	-
t_1998	2,795	5,05	-	-
t_1999	2,818	5,08	-	-
t_2000	2,810	5,10	-	-
t_2001	2,773	5,02	-	-
t_2002	2,768	5,01	-	-
t_2003	2,786	5,01	-	-
t_2004	2,799	5,00	-	-
ln_dochód	0,54	6,17	0,79	16,60
R^2	0,940	-	0,717	-
var $(\mu_i+\lambda_t)$	-	-	0,0004	-
var (v_{it})	0,0008	-	0,0010	-

Źródło: Obliczenia własne

Test Chowa potwierdził łączną istotność sztucznych zmiennych. Oznacza to, że model z dekompozycją wyrazu wolnego lepiej wyjaśnia kształtowanie się wydatków niż, oszacowany na tej samej próbie, model bez dekompozycji. Oceny wyrazów wolnych, dla kolejnych okresów badania, wyniosły od 15,92 – dla roku 2002 – do 16,74 w roku 1999. Oceny parametrów przy zmiennych sztucznych, wynikających z lokalizacji gospodarstwa przedstawiają relację wydatków konsumpcyjnych w

poszczególnych województwach do, przyjętych za podstawę, wydatków gospodarstw w województwie dolnośląskim. Ich wartości kształtowały się na poziomie od –8%, dla województwa mazursko-warmińskiego, do +4%, dla gospodarstw zamieszkałych w Mazowieckiem.

Statystyka mnożnika Lagrange'a dla modelu z dekompozycją składnika losowego (LM=106,9) przekroczyła wartość krytyczną z rozkładu χ^2 na poziomie istotności α=0,05. Oznacza to, że również model z efektami losowymi lepiej opisał wartość wydatków na ten cel od modelu pooled regression.

Modele z dekompozycją porównano przy pomocy testu Hausmana. Wartość empiryczna statystyki wyniosła H=10,69. Wynika z tego, że uwzględnione w modelu efekty grupowe i czasowe mają stały charakter.

7. Podsumowanie

Głównym celem artykułu było przedstawienie korzyści wynikających ze stosowania danych panelowych w modelowaniu ekonometrycznym. Omówiono budowę, metody estymacji i weryfikację modeli panelowych. Rozważania teoretyczne uzupełniono przykładem empirycznym, opisującym zróżnicowanie wydatków konsumpcyjnych gospodarstw domowych w Polsce. Hipotezę o ich przestrzennym i czasowym zróżnicowaniu zweryfikowano na podstawie danych statystycznych pozyskanych z Głównego Urzędu Statystycznego. W roli narzędzi badawczych wykorzystano jedno- i dwuczynnikowe modele panelowe ze sztucznymi zmiennymi oraz z dekompozycją składnika losowego.

Zastosowanie modeli ekonometrycznych pozwoliło ocenić siłę oddziaływania dochodów gospodarstw domowych na wysokość ponoszonych przez nie wydatków konsumpcyjnych. Dzięki wykorzystaniu modeli panelowych oszacowano różnice w wydatkach konsumpcyjnych

wynikające z lokalizacji gospodarstwa domowego. Ustalono także że mają one stały charakter. Oznacza to, że wysokość wydatków na towary i usługi konsumpcyjne gospodarstw domowych w Polsce zależy nie tylko od wysokości osiąganych przez nie dochodów. Jest ona także pochodną miejsca zamieszkania jego członków oraz czasu.

Literatura:

1. Baltagi B.H. 2001: Econometric Analysis of Panel Data. John Wiley & Sons Ltd., Chichester
2. Ciecieląg J., Tomaszewski A. 2003: Ekonometryczna analiza danych panelowych. WNE UW Warszawa
3. Dańska B. 1995: Wybrane metody estymacji modeli opartych na danych panelowych. UŁ Łódź
4. Grabiński T., Malina A., Zeliaś A. 1990: Metody analizy danych empirycznych na podstawie szeregów przekrojowo-czasowych. AE Kraków
5. Greene W. H. 2000: Econometric Analysis. Prentice-Hall, Inc., New Jersey
6. Welfe W. (red.) 1977: Ekonometryczne modele rynku, tom I: Metody ekonometryczne. PWE Warszawa
7. Suchecki B. (red.) 2000: Dane panelowe i modele wielowymiarowe w badaniach ekonomicznych, tom I: Dańska B. 2000: Przestrzenno-czasowe modelowanie zmian w działalności produkcyjnej w Polsce. Zastosowanie modeli panelowych. Absolwent Łódź
8. http://www.jmyc.republica.pl

***

Summary:

The aim of the paper was to present advantageous of panel data as well as panel data models. The methods of estimation and verification of panel data models were discussed as well. Theoretical discussion was accompanied by empirical example.

Adres do korespondencji:
dr Joanna Muszyńska
Wydział Nauk Ekonomicznych i Zarządzania UMK Toruń
Katedra Ekonometrii i Statystyki
ul. Gagarina 13a
87-100 Toruń
tel. (056) 6114784
e-mail: Joanna.Muszynska@uni.torun.pl