PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Barbara Dańska-Borsiak, Iwona Laskowska
Uniwersytet Łódzki

HETEROSKEDASTYCZNOŚĆ GRUPOWA I KORELACJA PRZEKROJOWA W ANALIZACH WYKORZYSTUJĄCYCH PRÓBY CZASOWO-PRZEKROJOWE

1. Wstęp

Dane czasowo-przekrojowe ze względu na swoją bogatą strukturę pozwalają na prowadzenie analiz w przekroju jednostek i czasu jednocześnie. Z tego powodu liczba praktycznych zastosowań modeli opartych na tego typu danych rośnie bardzo szybko. W analizie danych czasowo-przekrojowych pewnym zagadnieniom poświęca się więcej uwagi, innym mniej. Do często analizowanych zjawisk należą: heterogeniczność obiektów, autokorelacja w obrębie obiektu, heteroskedastyczność grupowa, różny rozwój obiektów w czasie, powodujący konieczność wprowadzania specyficznych dla obiektów zmiennych czasowych, niestacjonarność. Rzadziej analizowane są: autokorelacja przestrzenna, polegająca na skorelowaniu wartości jednej zmiennej dla różnych obiektów, i korelacja przekrojowa składnika losowego (jednoczesna, w tym samym okresie), występująca, jeśli skorelowane są ze sobą nieobserwowalne cechy różnych obiektów [Worrall, Pratt 2004].

Głównym celem artykułu jest zaprezentowanie metod testowania i uwzględniania w modelu heteroskedastyczności grupowej i korelacji przekrojowej oraz ich przykładowe zastosowanie do analizy popytu na pracę według sekcji PKD.

2. Rodzaje danych i przykłady zastosowań

Dane łączące wymiary przekrojowy i czasowy dzieli się na dwie podstawowe grupy: dane panelowe, charakteryzujące się duża liczbą obiektów n w stosunku do liczby okresów T oraz, w przeciwnym wypadku, dane czasowo-przekrojowe

(TSCS). Niektórzy badacze nazywają powyższe przypadki odpowiednio: sytuacją dominacji przekrojowej i sytuacją dominacji czasowej.

Modele oparte zarówno na danych panelowych, jak i TSCS wymagają zastosowania odpowiednich metod estymacji, bowiem przekrojowo-czasowa natura danych powoduje występowanie specyficznych problemów. W zależności od typu danych i wymienionych wyżej własności próby stosuje się różne techniki estymacyjne. W modelach (statycznych) szacowanych na podstawie danych panelowych zróżnicowanie przestrzenne określa się dwoma sposobami. Dla stosunkowo dużych wartości T możliwe jest wprowadzenie do modelu zmiennych sztucznych odpowiadających poszczególnym obiektom i stosowanie metod estymacji bazujących na MNK. Jeśli T jest bardzo małe (a n bardzo duże), to duża liczba zmiennych sztucznych może spowodować utratę zgodności estymatora. Wówczas różnice między obiektami odzwierciedla się przez różnicowanie części składnika losowego, a metody estymacji opierają się na UMNK.

Według niektórych autorów [Kristensen, Wawro 2003; Wilson, Butler 2004], podobne metody powinny być stosowane w przypadku danych TSCS. Większość badaczy zaleca jednak stosowanie wówczas innych procedur, ze względu na znaczne różnice w strukturze danych (typowe analizy panelowe prowadzone są zazwyczaj na próbach, w których T przybiera wartość nie przekraczającą 5, a dane TSCS zawierają z reguły co najmniej kilkanaście obserwacji w czasie). Różnice te sprawiają, że w estymacji modeli panelowych kluczowe znaczenie mają pewne specyficzne problemy, nie pojawiające się w analizach modeli TSCS. Najpopularniejszą metodą estymacji modeli opartych na danych TSCS, uwzględniającą zjawiska heteroskedastyczności grupowej i korelacji przekrojowej, jest UMNK z odpowiednio dobraną macierzą wariancji-kowariancji Ω.

Alternatywną metodą estymacji takich modeli jest zaproponowana przez N. Becka i L.N. Katza [Beck, Katz 1995; 1996] „regresja ze skorygowanymi dla danych panelowych błędami standardowymi" (ang. *panel corrected standard errors*), PCSE. Metoda ta krytykowana jest ze względu na dość skomplikowany sposób obliczania PCSE na podstawie reszt MNK modelu z opóźnioną zmienną objaśnianą [Worrall, Pratt 2004] oraz za to, że nie podejmuje się w niej próby modelowania procesów generujących występowanie zjawisk, a jedynie koryguje błędy standardowe [Maddala 1998].

W literaturze zarówno światowej, jak polskiej coraz częściej można spotkać analizy ekonomiczne wykorzystujące szeregi przekrojowo-czasowe. R.R. Kaufman i A. Segura-Ubiergo oparli badania efektu globalizacji, przejawiającego się zmianą wydatków na sferę socjalną zdrowie i edukację w 14 krajach Ameryki Łacińskiej, na danych czasowo-przestrzennych tych krajów w latach 1973-1997 [Kaufman, Segura-Ubiergo 2002]. Autorzy uwzględnili zjawisko heteroskedastyczności grupowej i korelacji przestrzennej (model oszacowano KMNK z PCSE).

L. Baccaro i D. Rei, wykorzystując dane przekrojowo-czasowe dla lat 1960--1998, przeprowadzili analizę wpływu szeregu czynników zarówno makroekono-

micznych, jak i instytucjonalnych na poziom bezrobocia w krajach OECD, uwzględniając takie problemy, jak niestacjonarność szeregów i estymacja panelowych modeli dynamicznych [Baccaro, Rei 2005].

W literaturze polskiej znaleźć można ciekawe przykłady analiz ekonomicznych wykorzystujących statyczne i dynamiczne modele panelowe, przykładem których może być praca K. Dudko-Kopczewskiej dotycząca efektów pierwotnej emisji akcji na Giełdzie Papierów Wartościowych [Dudko-Kopczewska 2004].

3. Testy statystyczne i metody estymacji

Metody analizy prowadzonej na podstawie „typowych" danych czasowo-przekrojowych (TSCS) są zbliżone do metod stosowanych w analizach szeregów czasowych, gdyż zbiory danych są z reguły na tyle długie, że umożliwiają prowadzenie analizy w kontekście procesów stochastycznych. Załóżmy, że dane pochodzą z n obiektów obserwowanych w T okresach, a zatem dysponujemy nT obserwacjami. Model estymowany na podstawie takich danych zapisać można w postaci ogólnej:

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{y} = \mathbf{X}\beta + \varepsilon \quad \Leftrightarrow \quad \mathrm{y}_{it} = x_{it}\beta + \varepsilon_{it}, \qquad (1)$$

gdzie dla $i = 1,...,n$ oraz $t = 1,...,T$:

$\mathbf{y}_i$ – wektor (T x 1) obserwacji na zmiennej objaśnianej, pochodzących z i-tego obiektu,

$\mathbf{X}_i$ – macierz (T x k+1) obserwacji na zmiennych objaśniających, pochodzących z i-tego obiektu,

β – wektor (k+1 x 1) parametrów strukturalnych,

ε_i – wektor (T x 1) składników losowych.

Specyfikacja (1) oznacza przede wszystkim przyjęcie założenia o jednakowych parametrach strukturalnych dla wszystkich obserwowanych obiektów. Naturalnie występujące pomiędzy obiektami różnice uwzględnia się poprzez przyjęcie odpowiednich założeń co do struktury macierzy wariancji – kowariancji składnika losowego. Podobnie uwzględnia się (oprócz połączenia danych w jeden szereg) powiązania pomiędzy nimi. Macierz wariancji – kowariancji składnika losowego równania (1) przybiera, w najbardziej ogólnym przypadku, postać:

$$\mathbf{V} = \begin{bmatrix} \sigma_{11}\Omega_{11} & \sigma_{12}\Omega_{12} & \cdots & \sigma_{1n}\Omega_{1n} \\ \sigma_{21}\Omega_{21} & \sigma_{22}\Omega_{22} & \cdots & \sigma_{2n}\Omega_{2n} \\ & \vdots & & \\ \sigma_{n1}\Omega_{n1} & \sigma_{n2}\Omega_{n2} & \cdots & \sigma_{nn}\Omega_{nn} \end{bmatrix}, \qquad (2)$$

przy czym $\mathbf{V}_{ij} = \sigma_{ij}\Omega_{ij}$ jest macierzą wariancji – kowariancji między składnikami losowymi pochodzącymi z *i*-tego i *j*-tego obiektu.

Heteroskedastyczność grupowa jest efektem tego, że różne obiekty podlegają różnym procesom (stochastycznym). Jej występowanie oznacza, że **V** jest macierzą diagonalną, a bloki na głównej przekątnej przybierają postać $\mathbf{V}_{ii} = \sigma_{ii}\mathbf{I}$. Do testowania hipotezy o występowaniu heteroskedastyczności grupowej stosować można szereg testów, m.in.: test White'a, test mnożnika Lagrange'a, test Walda. Pierwszy z nich wydaje się powszechnie znany. Statystyka empiryczna *testu mnożnika Lagrange'a* (H_0: $\sigma_{ii} = \sigma_{jj}$) przybiera w tym wypadku postać: $LM = \frac{T}{2}\sum_{i=1}^{n}(\frac{s_i^2}{s^2} - 1)^2$, gdzie s_i^2 jest oceną σ^2_{ii}, a s^2 wariancją obliczoną na podstawie reszt MNK dla całej próby (nT obserwacji). Statystyka LM ma rozkład χ^2 z $n - 1$ stopniami swobody.

Zjawisko *korelacji przekrojowej* polega na korelacji składników losowych pochodzących z różnych obiektów, ale z tego samego okresu. Jego występowanie wydaje się naturalne, gdyż łączenie danych przekrojowo-czasowych w jeden szereg jest zazwyczaj uwarunkowane tym, że analizowane obiekty są ze sobą powiązane, na przykład funkcjonują w obrębie jednego systemu gospodarczego (np. województwa), lub są elementami tej samej klasyfikacji (np. sekcje PKD). Wpływające na ich kondycję czynniki makroekonomiczne działają na nie w zróżnicowanym stopniu. Oznacza to, że każdy blok macierzy **V** przybiera postać:

$$\mathbf{V}_{ij} = \sigma_{ij}\mathbf{I}, \qquad i, j = 1,...,n . \tag{3}$$

Do testowania występowania zjawiska korelacji przekrojowej stosować można test mnożnika Lagrange'a (Breuscha-Pagana) lub test współczynnika wiarygodności.

Test Breuscha-Pagana wykorzystuje współczynniki korelacji reszt z *i*-tego i *j*-tego obiektu r_{ij}. Do weryfikacji hipotezy H_0: $\sigma_{ij} = 0$ dla $i \neq j$ służy statystyka $\lambda_{LM} = T\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij}^2$. Przy założeniu jednakowych parametrów strukturalnych dla wszystkich obiektów reszty te muszą być liczone na podstawie UMNK. Statystyka ta ma rozkład χ^2 z $n.(n - 1)/2$ stopniami swobody.

Test współczynnika wiarygodności weryfikuje występowanie korelacji *a posteriori*, gdyż wymaga znajomości ocen wariancji składnika losowego z modelu heteroskedastycznością grupową oraz modelu z heteroskedastycznością i korelacją przekrojową.

Najpopularniejszą metodą estymacji modeli uwzględniającą heteroskedastyczność grupową i korelację przekrojową jest UMNK z estymacją. Jej stosowanie jest w praktyce stosunkowo proste, gdyż odwracanie macierzy **V** nie nastręcza problemów. Zapisać ją bowiem można w postaci: $\mathbf{V} = \Sigma \otimes \mathbf{I}$, gdzie $\Sigma = \left[\sigma_{ij}\right]_T$, a zatem

$\mathbf{V}^{-1} = \Sigma^{-1} \otimes \mathbf{I}$. Zgodne estymatory elementów σ_{ij} otrzymuje się na podstawie reszt KMNK[1].

4. Wyniki analizy empirycznej

W celu zilustrowania wpływu heteroskedastyczności grupowej i korelacji przekrojowej na własności modelu oszacowano modele uwzględniające i nie uwzględniające tych zjawisk. W analizie wykorzystano dane statystyczne pochodzące z Badań Aktywności Ekonomicznej Ludności (BAEL)[2] dotyczące liczby pracujących w wybranych sekcjach, obejmujące okres od drugiego kwartału 1997 r. do trzeciego kwartału 2004 r. dla następujących sekcji PKD: budownictwo, handel, hotele i restauracje, przemysł, transport, administracja i ZUS, pośrednictwo finansowe, obsługa nieruchomości, edukacja, ochrona zdrowia, zaopatrywanie w energię elektryczną i gaz oraz sekcji pozostałe. W badaniach, ze względu na brak sprawozdawczości w tym okresie, pominięty został 2 i 3 kwartał roku 1999. Próba liczy 364 obserwacje, przy czym $T = 28$, $n = 13$.

W równaniach zatrudnienia oraz popytu na pracę w roli zmiennych niezależnych powinny wystąpić przynajmniej dwie zmienne: realny poziom produkcji i realna płaca przeciętna wyrażona w cenach producenta. Realny poziom produkcji reprezentować mogą wartość dodana brutto w cenach stałych, PKB w cenach stałych lub realna wartość produkcji sprzedanej. Płaca przeciętna może być wprowadzana bezpośrednio lub za pomocą indeksu Kaitza będącego ilorazem płacy minimalnej i płacy przeciętnej. Rolę tego współczynnika należy interpretować jako wpływ wzajemnych relacji między płacą minimalną a średnią, które oddziałują na strukturę zatrudnienia [Suchecki 2004; Dańska-Borsiak 2004]. Zmiany poziomu płacy minimalnej mogą powodować zmiany struktury popytu na pracę. Co więcej, wynikać one mogą ze zmian wszystkich rodzajów płac, a zatem mogą oddziaływać na zatrudnienie we wszystkich sekcjach.

Przyjęto założenie, że liczba pracujących w danej sekcji zależy od wartości dodanej wytworzonej w tej sekcji i relacji płacy minimalnej do płacy przeciętnej w danej sekcji (współczynnik Kaitza). Dodatkowo wprowadzone zostały zmienne sezonowe.

Analiza wpływu wymienionych czynników bazuje na następującym równaniu:

$$\ln(lprac)_{it} = \ln(\beta_0) + \beta_1 \ln(wdodana)_{it} + \beta_2 \ln(wsp.\ Kaitza)_{it} + \\ + \beta_3 * q1 + \beta_4 * q2 + \beta_5 * q3 + \xi_{it}, \tag{4}$$

[1] Rozważania prowadzono przy założeniu, że nie występuje autokorelacja składnika losowego w czasie. Możliwe jest rozszerzenie analizy o uwzględnienie również tego zjawiska [Greene 1997].

[2] Dane zaczerpnięte zostały z witryny internetowej www.sppp.gov.pl.

gdzie:

subskrypt *it* oznacza sekcję *i* w okresie *t*,

lprac – liczba pracujących (w tys. osób),

wdodana – wartość dodana brutto wytworzona w danej sekcji (c.s.1996)[3],

wsp. Kaitza – uproszczony współczynnik Kaitza liczony jako iloraz płacy minimalnej i płacy przeciętnej w danej sekcji,

$q1$, $q2$, $q3$ – zmienne zero-jedynkowe przyjmujące wartość 1 odpowiednio w I, II i III kwartale.

W powyższym modelu parametry są jednakowe dla wszystkich sekcji. Różnice między obiektami uwzględnione są poprzez własności składnika losowego. Wyniki estymacji prezentuje tab. 1.

Tabela 1. Oszacowania równania popytu na pracę

	KMNK	UMNK z uwzględnieniem heteroskedastyczności grupowej	UMNK z uwzględnieniem heteroskedastyczności grupowej i korelacji przekrojowej
Zmienna niezależna	*Oszacowane parametry*		
Wyraz wolny	8,253* (0,261)	8,291* (0,079)	8,239* (0,025)*
ln(wartość dodana)	0,392* (0,081)	0,257* (0,033)	0,267* (0,009)
ln(współczynnik Kaitza)	0,230** (0,331)	0,727* (0,137)	0,649* (0,036)
$q1$	-0,043*** (0,108)	0,045** (0,041)	0,043* (0,014)
$q2$	-0,074** (0,103)	-0,111* (0,035)	-0,090* (0,013)
$q3$	-0,042*** (0,104)	-0,108* (0,035)	-0,096* (0,013)
Test mnożnika Lagrange'a na heteroskedastyczność grupową	LM=273,61 $\chi^2(12)$=21,02		
Test mnożnika Lagrange'a na korelację przekrojową.		LM=949,00 $\chi^2(78)$=60,39	
Liczba obserwacji	364	364	364

W nawiasach pod ocenami parametrów podano wartości błędów średnich.

** – zmienne istotna na poziomie 5%,*

*** – zmienna istotna na poziomie 10%,*

**** – zmienna nieistotna statystycznie.*

Źródło: obliczenia własne w pakiecie Limdep 07.

Wartość dodana wytworzona w danej sekcji, zgodnie z oczekiwaniami, dodatnio wpływa na popyt na pracę. Również wzrost płacy minimalnej w stosunku do płacy przeciętnej powoduje wzrost popytu na pracę w poszczególnych sekcjach.

[3] W sprawozdawczości GUS wartość dodana przedstawiona jest z podziałem na sektor usług rynkowych i nierynkowych. Zatem dla sekcji występujących w obrębie danego sektora jest ona taka sama.

W pierwszym etapie zastosowania UMNK uwzględniona została heteroskedastyczność grupowa, pozostał zaś problem korelacji przekrojowej składnika losowego, na co wskazuje wartość statystyki LM = 949,00, znacznie przekraczająca wartość krytyczną χ^2. Świadczy to o istnieniu wzajemnych powiązań pomiędzy składnikami losowymi dla różnych sekcji w tym samym okresie, a tym samym o zależności wielkości zatrudnienia w danej sekcji od wielkości zatrudnienia w innych sekcjach.

Oceny parametrów strukturalnych dla etapów drugiego i trzeciego są bardzo podobne. Jednakże ostatni etap, uwzględniający obydwa omawiane problemy, przyniósł wzrost efektywności estymacji.

Uzyskane oceny parametrów strukturalnych informują o następujących zależnościach:

- jednoprocentowy wzrost wartości dodanej brutto powoduje, *ceteris paribus*, wzrost popytu na pracujących o 0,267%,
- jednoprocentowy wzrost stosunku płacy minimalnej do płacy przeciętnej powoduje, *ceteris paribus*, wzrost popytu na pracujących o 0,649%,
- średnia liczba pracujących w kwartale I jest o 43 osoby wyższa, w II kwartale – o 90 osób niższa, a w III kwartale o 96 osób niższa niż wynikająca z modelu wielkość w kwartale IV.

5. Podsumowanie

Przeprowadzone badania empiryczne pokazują, że uwzględnienie problemu heteroskedastyczności grupowej i korelacji przestrzennej przyczynia się do poprawy jakości modelu mierzonej wielkością błędów średnich ocen parametrów. W badaniu wykorzystano model ze stałymi parametrami strukturalnymi, odzwierciedlającymi średnie wartości dla wszystkich badanych sekcji. Stanowi to pewne ograniczenie w możliwościach analizy. Naturalnym rozszerzeniem rozważań byłoby zróżnicowanie parametrów względem sekcji, co będzie przedmiotem dalszych badań.

Literatura

Baccaro L., Rei D. (2005), *Isstitutional Determinants of Unemployment in OECD Countries: A Time-Series Cross-Section Analysis (1960-98)*, International Institute for Labour Studies, Discussion Paper, DP/160/2005.

Beck N., Katz L.N. (1995), *What to Do (and Not to Do) With Time-Series-Cross-Section Data*, „American Political Science Review" 89, s. 634-647.

Beck N., Katz L.N. (1996), *Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models*, „Political Analysis" 6, s. 1-34.

Dańska-Borsiak B. (2004), *Model struktury i prognoza liczby pracujących według wielkich grup zawodowych. Zastosowanie modelu o równaniach pozornie niezależnych*, [w:] *System prognozowania popytu na pracę w Polsce*, Studia i Materiały, CSRS, Warszawa, tom XII.

Dudko-Kopczewska K. (2004), *Analiza efektów pierwotnej emisji akcji na Giełdzie Papierów Wartościowych S.A. w Warszawie*, „Bank i Kredyt", luty 2004.

Greene W.H. (1997), *Econometric Analysis*, Prentice Hall International, Inc.

Kaufman R.R., Segura-Ubiergo A. (2002), *Globalization, Domestic Politics and Social Spendinig in Latin America: A Time-Series Cross-Section Analysis, 1973-1999*, http://econwpa.wustl.edu/eps/pe//papers/0504/0504009.pdf.

Kristensen I.P., Wawro G. (2003), *Lagging the Dog? The Robustness of Panel Corrected Standard Errors in the Presence of Serial Correlation and Observation Specific Effects.* Presented at the Annual Meeting of the Society for Political Methodology, University of Minnesota.

Maddala G.S. (1998), *Recent Developments in Dynamic Econometric Modeling: A Personal Viewpoint*; „Political Analysis" 7, s. 59-87.

Suchecki B. (2004), *Młodzież na rynku pracy w Polsce. Zastosowanie modeli dynamicznych i metod symulacyjnych do prognozowania popytu na pracę ludzi młodych według grup wieku*, RCSS, MZdPPP, Studia i Materiały, t. XII, s. 236-249.

Wilson S. E., Butler D.M. (2004), *A Lot More to Do: The Promise and Peril of Panel Data in Political Science* Manuscript, Department of Political Science, Brigham Young University.

Worrall J.L., Pratt T.C. (2004), *Estimation Issues with Time-Series – Cross-Section Analysis In Criminology*, „Western Criminology Review" 5(1), s. 35-49.

GROUPWISE HETEROSCEDASTICITY AND CROSS-SECTIONAL CORRELATION IN ANALYSES BASED ON TIME-SERIES – CROSS-SECTION DATA

Summary

Time-series – cross-section data (TSCS) are used more and more often in analyses of complex economic and social phenomena. Having combined the two dimensions – time and space results very often in groupwise heteroscedasticity and/or cross-sectional correlation. If those phenomena are taken into account while estimating models based on TSCS data, it is possible to increase the effectiveness of estimates and to model the processes which generate their appearance.

The methods of TSCS models testing and estimation are presented using a model of labour demand by PKD sections.