Analiza danych panelowych

Joanna Tyrowicz

18 kwietnia 2005r.

Dotychczas analizowaliśmy dwa rodzaje danych. Pierwszym z nich było obserwowanie wielu obiektów w tym samym czasie (np. gospodarstwa domowe, kraje, respondenci). Na tej podstawie wyciągaliśmy wnioski na temat charakteru jakiegoś zjawiska i jego uwarunkowań. Drugim rodzajem analizy było obserwowanie zmiennej (jednej lub kilku) w zmieniającym sie czasie (np. stopy procentowe, inflacja, stopy bezrobocia). Rodzaj relacji, jaki obserwowaliśmy pomiedzy tymi poszczegolnymi szeregami czasowymi określaliśmy mianem kointegracji.

Analiza panelowa pozwala połączyć te dwa rodzaje badań - w jednym badaniu określa sie zarówno wymiar podmiotów, jak i wymiar czasu. W tym materiale zrobimy to na podstawie autentycznych danych przygotowanych przez Jeffrey Frankel i Andrew K. Rose i opublikowanych pod tytułem *Estimating The Effect of Currency Union On Trade And Output* (NBER Working Paper No. 7875, 2002)[1].

Rodzaj analizy, który zostanie tu przeprowadzony nosi miano modelu grawitacyjnego i jest to technika bardzo popularna w badaniu handlu miedzynarodowego. Jej pionierem jest James Anderson, który w 1978 przedstawił teoretyczne podstawy tego podejścia. Ogólnie rzecz biorąc analizujemy model w którym z założenia:

$$Obroty\_Handlowe_{i,j,t} = \frac{PKB_{i,t} \cdot PKB_{j,t}}{Odlegl osc_{i,j}} \tag{1}$$

Jak widać z powyższego równania, mamy zarówno wymiar czasu (określony przez t) oraz obiektu (i, j). Zlogarytmowanie obu stron równania daje postać testowalną empirycznie:

$$obroty\_handlowe_{i,j,t} = \alpha + \beta_1(pkb_{i,t} + pkb_{j,t}) + \beta_2 odlegl osc_{i,j} + \epsilon_{i,j,t} \tag{2}$$

Ta postać testowalna została rozszerzona przez Frankel'a i Rose (2002) i nią bedziemy sie zajmować w dalszej cześci.

[1] Ścieżka dostepu do artykułu znajduje sie na stronie www.ekonometria.icm.edu.pl ⟶ Materiały

1 Dane

Zbiór zawiera dane dotyczące obrotów handlowych miedzy niemal wszystkimi krajami świata. Dane pochodzą z różnorodnych źródeł. Uporządkowane są w nastepujący sposób:

- Zmienną określającą rok dokonania obserwacji jest `year`. Obserwacje pochodzą z pieciolетnich przedziałów (od 1970 do 1995)
- W tych latach dokonano pomiarów dla niemal 8000 'par' krajów określonych w zbiorze jako `pairid`. Stanowi to próbke relacji handlowych około 180 krajów świata.
- Zmienną objaśnianą jest całkowity obrót handlowy w danym roku w obrebie danej pary (`ltrade`). Wielkości zostały zlogarytmowane, ponieważ tak podpowiada teoria.
- Zmienne objaśniające to po kolei:
  - Logartym sumy realnych PKB w obu krajach w danej parze w danym roku (`lrgdp`).
  - Logartym sumy populacji w obu krajach a danej parze w danym roku (`lpop`).
  - Logartym odległości miedzy dwoma krajmi w danej parze (`ldist`)
  - Zmienną zerojedynkową `cu` określającą przynależność dwóch danych krajów do jednego wspólnego obszaru walutowego (np. 12 krajów EMU).
  - Zmienną zerojedynkową `comlang` określającą, czy dane dwa kraje posługują sie tym samym jezykiem, jako jezykiem oficjalnym (np. Belgia ma jezyk wspolny z krajami francuskojezycznymi, flamandzkojezycznymi oraz niemieckojezycznymi).
  - Zmienną zerojedynkową `border` określającą, czy dane dwa kraje mają wspólną granice lądową (np. Polska z Niemcami, Czechami, Słowacją, Ukrainą, Białorusią, Litwą i Rosją).
  - Zmienną bedącą sumą dwóch zmiennych zerojedynkowych `colonial` określającą, czy dane dwa kraje miały w ogóle przeszłość kolonialną (np. Polska i Austria razem nie)
  - Zmienną zerojedynkową `comcol` określającą, czy dane dwa kraje miały w przeszłości wspólnego kolonizatora (np. Australia i Kanada tak, a Australia i Japonia nie).
  - Zmienną zerojedynkową `comctry` o ile dane dwa kraje tworzyły kiedyś jeden kraj (np. Ghana i Rwanda).

– Zmienną zerojedynkową `regional` o ile dane dwa kraje należą do jednego wspólnego ugrupowania wspierającego handel miedzy jego członkami (np. NAFTA, ASEAN, EFTA)

– Zmienną `island` bedącą sumą zmiennych zerojedynkowych, jeśli dwa kraje są wyspami.

– Zmienną `ll` bedącą sumą zmiennych zerojedynkowych, jeśli dwa kraje nie mają w ogóle dostepu do morza.

2 Praca w STATA

Zanim otworzycie państwo plik `panel_data.dta`, należy uruchomić program STATA i wpisać w nim nastepujące komendy:

```
set memory 99m
set matsize 800
```

W ten sposób ustawiliśmy w programie maksymalne 'moce przerobowe' - plik na którym bedziemy operować jest bardzo duży i bez tych ustawień nie jest możliwe nawet jego otwarcie. Jeśli zrobićby to w odwrotnej kolejności (najpierw otworzyć zbiór danych a potem ustawić parametry) STATA na to nie pozwoli.

Przy użyciu komendy `describe` można obejrzeć opisy zmiennych przygotowane przez autorów artykułu. Są to opisy tekstowe lepiej wyjaśniające nature każdej zmiennej.

2.1 Regresja standardowa a badania panelowe

Przy użyciu komendy `reg` lub `regress` możemy uzyskać na tym zbiorze wyniki standardowej regresji, która nie uwzglednia faktu, że pracujemy z danymi panelowymi.

```
. regress ltrade lrgdp lpop ldist
```

Sprawdźmy teraz na ile wiarygodne są te wyniki. Aby przeprowadzić regresje panelową, trzeba określić zmienne służące do zdefiniowania wymiarów. Wymiar czasu określa sie komendą `tis` a wymiar obiektu komenda `iis`.

```
. tis year
. iis pairid
```

Należy przy tym zaznaczyć, że to od nas zalezy jak okreslimy te wymiary.

Z punktu widzenia analizy, komenda `iis` określa zasade grupowania zmiennych. Przy takim zdefiniowaniu jak powyżej, STATA potworzy dla każdej pary krajów grupe z taką liczbą ob-

serwacji, dla ilu lat mamy dla tej pary dane. Gdyby zrobić odwrotnie, STATA stworzyłaby grupe w danym roku dla wszystkich dostepnych par krajów. Nie ma zasady, która mówi, że należy postepować tak, a nie inaczej. Trzeba mieć jednak świadomość, że to zdefiniowanie bedzie miało wpływ na nasze wyniki - zdefiniowanie tych kryteriów grupowania określa rodzaj odpowiedzi, który dostaniemy w wyniku regresji panelowej. Określiwszy wymiary, możemy wykonać regresje panelową. Robimy to komendą:

```
. xtreg ltrade lrgdp lpop ldist
```

Najpierw porównajmy wyniki obu estymacji. Jak widać, przedziały ufności nawet sie na siebie nie na chodza (poza zmienna `ldist`). Oznacza to, ze estymacja danych panelowych przy uzyciu standardowej regresji jest po prostu niepoprawna.

2.2 Interpretacja analizy panelowej

Aby zinterpretować wyniki regresji panelowej, wykonajmy całą regresje zaproponowaną przez Frankela i Rose'a[2].

```
. xtreg ltrade lrgdp lpop ldist cu comlang border  comcol comctry colonial ll regional
```

Interpretując po kolei wyniki estymacji należy zwrócić uwage na tytuł nadany przez STATA tej analizie, czyli stwierdzenie `Random-effects GLS regression`. Określenie `Random-effects` wynika z tego, iż STATA regresje z efektami zmiennymi wykonuje domyślnie. Gdybyśmy chcieli wykonać regresje z efektami stałymi (*fixed effect*) należy powyższą komedne zmodyfikować dodając na jej końcu opcje `, fe`[3].

Prawy panel podsumowania wyników podaje nam liczbe grup (7 961) oraz obserwacji (31 226). Podaje również liczbe obserwacji w grupie: minimalną (1, inaczej nie byloby grupy), maksymalną (6, mamy tylko 25 lat piecioletnich obserwacji) oraz średnią (3.9). Na końcu podaje wyniki testu na łączną nieistotność parametrów (statystyka Walda).

Lewy panel raportuje która ze zmiennych została określona za grupującą. Po drugie raportuje wielkości R^2. Dostaliśmy R^2 *within*, R^2 *between* oraz R^2 *overall*. Każda z tych statystyk ma inną interpretacje. Jeśli chodzi o R^2 *within* związane ono jest z wariancją wewnątrz grupy. W naszym przypadku grupy są bardzo małe i nic dziwnego, że dopasowanie w tym zakresie jest niewielkie. Statystyka R^2 *between* opisuje relacje wariancji w przestrzeni pomiedzy grupami. W przypadku tych badań, to właśnie R^2 *between* jest tą statystyką, która nas interesuje najbardziej. Odpowiada nam ona na pytanie, jak dobrze przy użyciu modelu grawita-

[2]Zmienna `island` jest nieistotna, wiec nie bedziemy jej dalej rozpatrywac.

[3]Do kwestii wyboru efektów stałych czy zmiennych jeszcze wrócimy

cyjnego umiemy wyjaśnić schematy w handlu miedzynarodowym i okazuje sie, że w około 63%. Statystyka R^2 *overall* jest ważonym uśrednieniem tych dwóch statystyk i nie bardzo poddaje sie interpretacji innej niż intuicyjna.

Poniżej statystk R^2 znajduje sie informacja o tym, ze założono rozkład normalny dla efektów zmiennych oraz, że założono zerową korelacje miedzy tymi efektami a macieżą zmiennych objaśniających. Wyjaśnieniu tej kwestii poświecony jest nastepny podrozdział.

2.3 Efekty stałe, czy zmienne?

Wybór efektów stałych lub zmiennych może być podyktowany przez teorie, co nie wyklucza przetestowania słuszności tego wyboru na danym zbiorze danych. Wracając do równania (2) ze wstepu do tych materiałów, kwestie efektów w modelu można przedstawić jako wybór miedzy nastepującymi alternatywami:

$$obroty\_handlowe_{i,j,t} = \alpha + \beta_1(pkb_{i,t} + pkb_{j,t}) + \beta_2 odleglosc_{i,j} + \epsilon_{i,j,t} \tag{3}$$

oraz

$$obroty\_handlowe_{i,j,t} = \alpha_{i,j} + \beta_1(pkb_{i,t} + pkb_{j,t}) + \beta_2 odleglosc_{i,j} + \epsilon_{i,j,t} \tag{4}$$

Dodanie indeksów do parametru α równoznaczne jest z zezwoleniem na efekty zmienne. Intuicyjnie, oznacza to, iż pozwalamy by handel autonomiczny (czyli niewyjaśniany żadną ze zmiennych w modelu) był inny dla każdej pary krajów. Wybór efektów stałych oznacza iż uważamy, że jest on taki sam dla pary Ghana-Polska jak dla pary Holandia-Belgia. Choć już na pierwszy rzut oka wydaje sie, że w tym modelu powinniśmy zezwolić na efekty zmienne, aby mieć pewność słuszności wyboru, powinniśmy przeprowadzić testy weryfikujące.

Do określenia tego możemy skorzystać z dwóch rodzajów testów wbudowanych w STATA. Oba mają podobną interpretacje choć nieco inne konstrukcje.

Test Breuscha-Pagana skonstruowany na bazie mnożników Lagrange bada, czy wariancja wynikająca z par (a nie z wymiaru czasu) jest statystycznie istotna. Hipoteza zerowa mówi o tym, iż wariancja ta nie odgrywa żadnej roli (prawdziwy jest model efektów stałych). Wywołuje go komenta `xttest0`. W tym przypadku test bardzo silnie odrzuca hipoteze zerową na rzecz alternatywnej.

Test Hausmana estymuje model z efektami stałymi, drugi z efektami zmiennymi i porównuje ze sobą wyestymowane współczynniki. Jeśli są od siebie różne w sposób statystycznie istotny,

to znaczy, że założenie o efektach stałych nie było słuszne. Test Hausmana wywołuje komenda `xthausman`.

Obie komendy można wywołać jedynie po regresji typu RE (efekty zmienne). Jeśli bedziemy chcieli skorzystac z tych testow po regresji typu FE (efekty stałe), STATA nie poda wyników.

Należy zaznaczyć, że jeżeli w danych naprawde wystepują efekty stałe, estymator efektów zmiennych (domyślny w STATA) jest efektywny i zgodny. Natomiast odwrotne stwierdzenie nie jest prawdziwe: narzucenie założenia o efektach stałych na modelu, gdzie w rzeczywistości wystepują efekty zmienne, powoduje otrzymanie niewłaściwych estymatorów.

Wracając do informacji o Gausowskim rozkładzie czynnika stałego, w przypadku estymacji typu RE (efekty zmienne) STATA tak naprawde estymuje dwie stale w modelu. Po pierwsze stałą 'wspólną' dla wszystkich zaraportowaną jako 'constant' wśród zmiennych objaśniających. Drugą jest ten specyficzny dla danego obiektu (u nas pary krajów) efekt stały.

Stwierdzenie o założonym rozkładzie Gausowskim jest równoważne temu, że STATA przyjeła iż efekty stałe mają jakiś rozkład i jest to rozkład normalny. Jest to potrzebne tylko i wyłącznie do drugiej cześci tego stwierdzenia, a mianowicie: `corr(u_i, X) = 0 (assumed)`. Oznacza to, iż mimo iż efekty stałe mają rozkład normalny, a w dużych próbach (takich jak nasza) macież X również powinna mieć rozkład zbliżony do normalnego, STATA nie dopuszcza równoczesnej korelacji miedzy efektami stałymi a macieżą X[4]

Efekty tego obserwujemy również w statystykach znajdujących sie poniżej estymatorów parametrów. Mamy tam podane wartości σ_u, czyli wariancji wynikającej z różnorodności miedzy parami oraz σ_e, czyli wariancji wynikającej z wymiaru czasu. Współczynnik ρ mówi nam o tym, jaka cześć wariancji (błedów standardowych) w modelu wyjaśniona jest przez efekty właściwe dla pary. Jeśli ρ jest relatywnie duże w naszym przypadku, to znaczy, że model opisuje zjawisko dość stabilne w czasie. A to oznacza, że nie musimy sie szczególnie martwić niską wartością $R^2 within$ ponieważ proces który analizujemy jest dość dobrze wyjaśniony.

2.4 Efekty stałe a efekt roku?

Do tej pory analizowaliśmy cały czas jakiś specyficzny efekt stały związany z obserwowaniem konkretnej pary krajów. Nie sprawdzaliśmy natomiast, czy przypadkiem nie wystepuje również stały i specyficzny efekt związany z momentem obserwacji. STATA jest bardzo dobrze wyposażona

[4]Jest to założenie równoważne założeniu ortogonalności reszt w modelach MNK. Nie testuje sie go - jeśli teoria podpowiada endogeniczność którejś ze zmiennych macieży X, estymatory mogą nie być zgodne, a jedynym rozwiązaniem jest użycie instrumentów.

do analizowania tego typu efektów - korzystając z komendy `xi` możemy zarządać zweryfikowania hipotezy o istotności tego typu efektów.

```
xi:xtreg ltrade lrgdp lpop ldist cu ... i.year
```

Jak widać, estymatory przy wszystkich latach są istotne, wiec pominiecie tego efektu mogło prowadzić do obciążoności estymatorów w poprzednich badaniach. Warto również zwrócić uwage iż regresja

```
xi:xtreg ltrade lrgdp lpop ldist cu ... i.pairid
```

jest równoznaczna z wyestymowaniem *explicite* wszystkich efektów stałych w modelu. Prosze jednak spróbować to zrobić - w wiekszości przypadków po dłuższej chwili namysłu STATA poda komunikat iż nie jest w stanie przeprowadzić tej operacji ze wzgledu na ograniczenia pamieci.

Jak zauważycie w literaturze, w przypadku analiz danych panelowych bardzo czesto uwzglednia sie ten element w regresji, natomiast nieczesto raportuje w wynikach badań (widać to na przykład w uwagach pod tabelami w artykule Frankela i Rose'a).