

# ANALIZA DANYCH PANELOWYCH

Joanna Tyrowicz

14 kwietnia 2007r.

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 się czasie (np. stopy procentowe, inflacja, stopy bezrobocia). Rodzaj relacji, jaki obserwowaliśmy pomiędzy tymi poszczególnymi szeregami czasowymi określaliśmy mianem kointegracji.

Analiza panelowa pozwala połączyć te dwa rodzaje badań - w jednym badaniu określa się 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)<sup>1</sup>.

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

$$Obroty\_Handlowe_{i,j,t} = \frac{PKB_{i,t} \cdot PKB_{j,t}}{Odleglosc_{i,j}} \quad (1)$$

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

$$obroty\_handlowe_{i,j,t} = \alpha + \beta_1(pk_{i,t} + pk_{j,t}) + \beta_2 odleglosc_{i,j} + \epsilon_{i,j,t} \quad (2)$$

Ta postać testowalna została rozszerzona przez Frankel'a i Rose (2002) i nią będziemy się zajmować w dalszej części.

---

<sup>1</sup>Ścieżka dostępu do artykułu znajduje się na stronie [www.ekonometria.icm.edu.pl](http://www.ekonometria.icm.edu.pl) → Materiały

# 1 Dane

Zbiór zawiera dane dotyczące obrotów handlowych między niemal wszystkimi krajami świata.

Dane pochodzą z różnorodnych źródeł. Uporządkowane są w następujący sposób:

- Zmienną określającą rok dokonania obserwacji jest `year`. Obserwacje pochodzą z pięcioletnich przedziałów (od 1970 do 1995)
- W tych latach dokonano pomiarów dla niemal 8000 'par' krajów określonych w zbiorze jako `pairid`. Stanowi to próbkę relacji handlowych około 180 krajów świata.
- Zmienną objaśnianą jest całkowity obrót handlowy w danym roku w obrębie danej pary (`ltrade`). Wielkości zostały zlogarytmowane, ponieważ tak podpowiada teoria.
- Zmienne objaśniające to po kolei:
  - Logarytm sumy realnych PKB w obu krajach w danej parze w danym roku (`lrgdp`).
  - Logarytm sumy populacji w obu krajach a danej parze w danym roku (`lpop`).
  - Logarytm odległości między dwoma krajami 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 się tym samym językiem, jako językiem oficjalnym (np. Belgia ma język wspólny z krajami francuskojęzycznymi, flamandzkojęzycznymi oraz niemieckojęzycznymi).
  - Zmienną zerojedynkową `border` określającą, czy dane dwa kraje mają wspólną granicę lądową (np. Polska z Niemcami, Czechami, Słowacją, Ukrainą, Białorusią, Litwą i Rosją).
  - Zmienną będącą 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 między jego członkami (np. NAFTA, ASEAN, EFTA)
- Zmienną `island` będącą sumą zmiennych zerojedynkowych, jeśli dwa kraje są wyspami.
- Zmienną `ll` będącą sumą zmiennych zerojedynkowych, jeśli dwa kraje nie mają w ogóle dostępu do morza.

## 2 Praca w STATA

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

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

W ten sposób ustawiliśmy w programie maksymalne 'moce przerobowe' - plik na którym będziemy operować jest bardzo duży i bez tych ustawień nie jest możliwe nawet jego otwarcie. Jeśli zrobiłoby 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 naturę 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 uwzględnia faktu, że pracujemy z danymi panelowymi.

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

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

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

Należy przy tym zaznaczyć, że to od nas zależy jak określimy te wymiary.

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

serwacji, dla ilu lat mamy dla tej pary dane. Gdyby zrobić odwrotnie, STATA stworzyłaby grupę w danym roku dla wszystkich dostępnych par krajów. Nie ma zasady, która mówi, że należy postępować tak, a nie inaczej. Trzeba mieć jednak świadomość, że to zdefiniowanie będzie miało wpływ na nasze wyniki - zdefiniowanie tych kryteriów grupowania określa rodzaj odpowiedzi, który dostaniemy w wyniku regresji panelowej.

Określiwszy wymiary, możemy wykonać regresję panelową. Robimy to komendą:

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

Najpierw porównajmy wyniki obu estymacji. Jak widać, przedziały ufności nawet się na siebie nie nachodzą (poza zmienną `ldist`). Oznacza to, że estymacja danych panelowych przy użyciu standardowej regresji jest po prostu niepoprawna.

## 2.2 Interpretacja analizy panelowej

Aby zinterpretować wyniki regresji panelowej, wykonajmy całą regresję zaproponowaną przez Frankela i Rose'a<sup>2</sup>.

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

Interpretując po kolei wyniki estymacji należy zwrócić uwagę na tytuł nadany przez STATA tej analizie, czyli stwierdzenie `Random-effects GLS regression`. Określenie `Random-effects` wynika z tego, iż STATA regresję z efektami zmiennymi wykonuje domyślnie. Gdybyśmy chcieli wykonać regresję z efektami stałymi (*fixed effect*) należy powyższą komendę zmodyfikować dodając na jej końcu opcję `, fe`<sup>3</sup>.

Prawy panel podsumowania wyników podaje nam liczbę grup (7 961) oraz obserwacji (31 226). Podaje również liczbę obserwacji w grupie: minimalną (1, inaczej nie byłoby grupy), maksymalną (6, mamy tylko 25 lat obserwacji, które grupujemy sobie w okresy pięcioletnie, aby pozbyć się efektów związanych z oczywistą cyklicznością handlu i PKB) oraz średnią liczbę obserwacji w grupie (3.9). Na końcu podaje wyniki testu na łączną nieistotność parametrów (statystyka Walda).

Lewy panel po pierwsze 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 interpretację. 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

---

<sup>2</sup>Zmienna `island` jest nieistotna, więc nie będziemy jej dalej rozpatrywać.

<sup>3</sup>Do kwestii wyboru efektów stałych czy zmiennych jeszcze wrócimy

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

Poniżej statystyk  $R^2$  znajduje się informacja o tym, że założono rozkład normalny dla efektów zmiennych oraz, że założono zerową korelację między tymi efektami a macierzą zmiennych objaśniających. Wyjaśnieniu tej kwestii poświęcony jest następny podrozdział.

### 2.3 Efekty stałe, czy zmienne?

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

$$\text{obroty\_handlowe}_{i,j,t} = \alpha + \beta_1(\text{pkb}_{i,t} + \text{pkb}_{j,t}) + \beta_2\text{odleglosc}_{i,j} + \epsilon_{i,j,t} \quad (3)$$

oraz

$$\text{obroty\_handlowe}_{i,j,t} = \alpha_{i,j} + \beta_1(\text{pkb}_{i,t} + \text{pkb}_{j,t}) + \beta_2\text{odleglosc}_{i,j} + \epsilon_{i,j,t} \quad (4)$$

Dodanie indeksów do parametru  $\alpha$  równoznaczne jest z zezwoleniem na efekty nieobserwowalne (indywidualne). Intuicyjnie, oznacza to, iż pozwalamy by handel autonomiczny, to jest część handlu, która nie zmienia się w czasie była różna dla poszczególnych par krajów. Jeśli używamy estymatora efektów losowych, że to oznaczać będzie, że zakładamy, że handel autonomiczny nie jest skorelowane ze zmiennymi objaśniającymi. Wybór estymatora efektów stałych oznacza iż uważamy, że handel autonomiczny może być skorelowany ze zmiennymi objaśniającymi. Oznacza zatem iż uważamy, że jest on taki sam dla pary Ghana-Polska jak dla pary Holandia-Belgia.

Choć już na pierwszy rzut oka wydaje się, 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 na podstawie wyników testów weryfikujących założenie o braku korelacji między efektami indywidualnymi a zmiennymi objaśniającymi. Oba mają podobną interpretację 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. Test Breuscha-Pagana stosujemy w przypadku estymatora efektów losowych. Skonstruowany on jest na bazie mnożników Lagrange i 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 (a więc prawdziwy jest model, w którym efekty nieobserwowalne nie występują). Hipoteza zerowa mówi o tym, iż wariancja ta nie odgrywa żadnej roli (prawdziwy jest model efektów stałych). Podobną hipotezę hipotezę w przypadku estymatora efektów stałych badamy testując, czy wszystkie  $\alpha$  są łącznie równe zeru (statystyka F). Wywołuje go komenda `xttest0`. W tym przypadku test bardzo silnie odrzuca hipotezę 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 będziemy chcieli skorzystać z tych testów po regresji typu FE (efekty stałe), STATA nie poda wyników. Należy zaznaczyć, że jeżeli w danych naprawdę występują efekty losowych (nie ma korelacji między efektami nieobserwowalnymi a zmiennymi objaśniającymi), to estymator efektów stałych (domyslny w STATA) jest zgodny. Natomiast odwrotne stwierdzenie nie jest prawdziwe: przyjęcie założenia o efektach losowych w modelu, w którym efekty nieobserwowalne są skorelowane ze zmiennymi objaśniającymi spowoduje otrzymanie estymatora, który nie jest zgodny.

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

Ponizej estymatorów parametrów, mamy także podane wartości odchylenia standardowego dla  $u$ , wynikającego ze zróżnicowania efektów nieobserwowalnych między parami oraz odchylenie standardowe dla  $e$ , związane z błędem losowym, który nie jest stały w czasie. Współczynnik  $\rho$  mówi nam o tym, jaka część wariancji całkowitego błędu losowego (sumy efektu indywidualnego i czystego błędu losowego) w modelu wyjaśniona jest przez efekty właściwe dla pary.

## 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 występuje również stały i specyficzny efekt związany z momentem obserwacji. STATA jest bardzo dobrze wyposażona do analizowania tego typu efektów - korzystając z komendy `xi` możemy zarządzić zweryfikowania hipotezy o istotności tego typu efektów.

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

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

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

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

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