

Mikroekonometria

11

Mikołaj Czajkowski
Wiktor Budziński

Dane panelowe

- ▶ Co jeśli mamy do dyspozycji dane panelowe?
 - ▶ Kilka obserwacji od tych samych respondentów, w różnych punktach czasu (np. ankieta realizowana co roku)
 - ▶ Możemy spodziewać się korelacji składników losowych (ε) dla tych samych osób (nawet jeśli pomiędzy osobami nadal będą niezależne)
 - ▶ Połączenie wszystkich danych i zignorowanie korelacji powoduje problemy (estymator ML nie będzie zgodny)
 - ▶ Estymacja wszystkich korelacji wymaga szacowania prawdopodobieństwa z T_i -mianowego rozkładu
 - ▶ Trudne z uwagi na oceniany charakter obserwowalnej zmiennej (np. 0 lub 1) i zwykle niewielką liczbę obserwowanych okresów dla osoby (T_i)



Dane panelowe – modele binarne

► Specyfikacja dla danych przekrojowych

$$\begin{aligned} Y_{it}^* &= \mathbf{X}_{it}\boldsymbol{\beta} + \varepsilon_{it} \\ Y_{it} &= \mathbf{1}(\mathbf{X}_{it}\boldsymbol{\beta} + \varepsilon_{it} > 0); \quad i=1,\dots,N; \quad t=1,\dots,T_i \end{aligned}$$

► Specyfikacja dla danych panelowych

$$\begin{aligned} Y_{it}^* &= \mathbf{X}_{it}\boldsymbol{\beta} + \nu_{it} + u_i \\ Y_{it} &= \mathbf{1}(\mathbf{X}_{it}\boldsymbol{\beta} + \nu_{it} + u_i > 0); \quad i=1,\dots,N; \quad t=1,\dots,T_i \end{aligned}$$

- u_i mierzy nieobserwowalną, specyficzną dla osoby heterogeniczność
- Forma zależności między u_i a $\mathbf{X}_{it}$ daje model z efektami 'losowymi' lub 'stałymi'

Dane panelowe – modele binarne

► Specyfikacja dla danych panelowych

$$Y_{it}^* = \mathbf{X}_{it}\boldsymbol{\beta} + \nu_{it} + u_i$$
$$Y_{it} = \mathbf{1}(Y_{it}^* > 0) \quad ; \quad i=1,\dots,N; \quad t=1,\dots,T_i$$

► Model z efektami stałymi – u_i i $\mathbf{X}_{it}$ mogą być skorelowane

- u_i to jakiś element ε_{it} , który jest stały dla danego respondenta w każdym okresie
- Brak restrykcji, ale może powodować 'problem przypadkowych parametrów'

► Model z efektami losowymi – u_i i $\mathbf{X}_{it}$ są niezależne

- Zakładamy, że u_i ma jakiś (określony) rozkład w populacji
 - Narzuca restrykcje na rozkład heterogeniczności w populacji
- Forma modelu nie oznacza, że efekty są takie czy inne – to tylko wybór ekonometrycznego podejścia do problemu

Zadanie 1. Opieka zdrowotna w Niemczech

- ▶ Wykorzystując projekt `me.gerhealth.dta` skonstruuj model, w którym odbycie wizyty u lekarza ($Y = \mathbf{1}(\text{docvis} > 0)$) wyjaśniane jest przez stałą, płeć, wiek, dochód, posiadanie dzieci, liczbę lat edukacji i bycie w małżeństwie
- 1. Potraktuj wszystkie obserwacje jako niezależne
- 2. Skonstruuj model efektów losowych
 - 1. Czy korelacje są istotne?
 - 2. Dlaczego model konverguje dłużej?
 - 3. Porównaj wyniki (charakterystyki modeli, parametry, efekty krańcowe)



Dane panelowe – modele binarne

- ▶ Model efektów losowych nakłada pewne restrykcje

- ▶ Określony rozkład u_i
- ▶ Stałe korelacje pomiędzy okresami

- ▶ Model efektów stałych

$$\begin{aligned} Y_{it}^* &= \alpha_i d_{it} + \mathbf{X}_{it} \boldsymbol{\beta} + \varepsilon_{it} \\ Y_{it} &= \mathbf{1}(Y_{it}^* > 0) \end{aligned} ; \quad i=1, \dots, N; \quad t=1, \dots, T_i$$

- ▶ $d_{it} = 1$ dla respondenta i , 0 w p.p. (w modelu jedna stała dla każdej osoby)
- ▶ $\mathbf{X}_{it}$ nie zawiera teraz stałej
- ▶ Model wymaga estymacji dodatkowych $N-1$ parametrów!
 - ▶ Mogą ich być tysiące

Dane panelowe – modele binarne

Model efektów stałych

► Funkcja LL dla modelu efektów stałych

$$\ln L = \sum_{i=1}^N \sum_{t=1}^T \ln \Pr(Y_{it} | \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta})$$

- Krańcowe prawdopodobieństwo (Pr) może mieć jedną z form poznanych wcześniej (np. logit, probit, ...)
- Funkcja nieco trudniejsza w estymacji (z powodu rozmiaru), ale nie to jest głównym problemem
- Problemy statystyczne

Dane panelowe – modele binarne

Model efektów stałych

► Możliwe problemy statystyczne

- W praktyce – zwykle T_i jest stałe i ... małe
- Każde α_i szacowane jest na podstawie T_i obserwacji
- Model może nie konwergować
- Estymator β także może nie być zgodny, bo jest funkcją α
- To jest tzw. 'problem przypadkowych parametrów'
 - Obciążenie rzędu 100% ($T = 2$) – 10% ($T = 8$)
- Model jest nieidentyfikowalny jeśli występują w nim zmienne objaśniające, które nigdy nie zmieniają się z okresu na okres
- Model jest nieidentyfikowalny dla obserwacji, dla których zmienna objaśniana ma tę samą wartość w każdym okresie (m.in. dla obserwacji dla których $T_i = 1$)

Zadanie 1. Opieka zdrowotna w Niemczech c.d.

- ▶ Wykorzystując projekt `me.gerhealth.dta` skonstruuj model, w którym odbycie wizyty u lekarza ($Y = \mathbf{1}(\text{docvis} > 0)$) wyjaśniane jest przez stałą, płeć, wiek, dochód, posiadanie dzieci, liczbę lat edukacji i bycie w małżeństwie
- 3. Skonstruuj model efektów stałych
 - 1. Czy wszystkie obserwacje zostały wykorzystane?
 - 2. Czy wszystkie zmienne zostały wykorzystane?
 - 3. Który model lepiej pasuje do danych RE czy FE?
- 4. Dodaj efekty stałe dla lat



Dane Panelowe

- ▶ Analogicznie, panelowy charakter mogą mieć dane uporządkowane i o liczności zdarzeń

- ▶ `xtpoisson` – panelowa regresja Poissona
- ▶ `xtnbreg` – panelowa regresja ujemna dwumianowa
- ▶ `xtologit` – panelowa regresja uporządkowana

Praca domowa ME.11 (grupy 3-osobowe)

1. Dane zawarte w `me.usahealth.dta` pochodzą z eksperymentu społecznego w USA, w którym rodziny losowo przyporządkowano do różnych rodzajów ubezpieczeń medycznych i przez kilka lat zbierano dane o tym, jak z nich korzystają. Celem badania było sprawdzenie, w jaki sposób wysokość udziału własnego (*coinsurance rate*) wpływa na korzystanie z usług medycznych. Proszę to zweryfikować, kontrolując inne istotne zmienne.
 1. Uwzględniając panelowy charakter danych, skonstruuj model efektów losowych (liczności zdarzeń) wyjaśniający to, ile razy osoba poszła do lekarza (*mdu*). Porównaj model Poissona z modelem Ujemnym dwumianowym.
 2. Uwzględniając panelowy charakter danych, skonstruuj model efektów losowych wyjaśniający to, czy osoba poniosła jakiekolwiek wydatki na pozaszpitalną opiekę medyczną (*med* > 0).
 3. W powyższych modelach zinterpretuj wyniki jakościowo oraz ilościowo
 4. Zbadaj, czy w powyższych modelach istnieją statystycznie istotne efekty stałe dla lat. Czy w poszczególnych latach liczba wizyt u lekarza lub p-stwo poniesienia kosztu na opiekę medyczną były takie same?
 5. Czy cel badania można osiągnąć stosując modele z efektami stałymi?

```
set seed 10+"Nr indeksu"  
sample2 90, cluster(id)
```



Model parametrów losowych

- ▶ W modelu efektów losowych zakładamy, że stała może być różna dla różnych respondentów i że indywidualne stałe przyjmują wartości się według określonego rozkładu
- ▶ Teoretycznie nie ma powodów, żeby ograniczać się tylko do stałej
 - ▶ Może inne parametry również powinny być losowe?
 - ▶ Widzieliśmy podczas wcześniejszych zajęć przykłady obserwowalnej heterogeniczności preferencji
 - ▶ Co jeśli ta heterogeniczność nie jest powodowana zmiennymi, które mamy w próbie?
 - ▶ Jak uwzględnić nieobserwowaną heterogeniczność?



Model parametrów losowych

▶ Wróćmy do wyboru wielomianowego

$$U_{ijt} = \mathbf{x}_{ijt} \boldsymbol{\beta}_i + \varepsilon_{ijt}$$

- ▶ Dane panelowe – indeks t mówi o sytuacji wyboru tego samego respondenta
- ▶ Parametry są indeksowane po i , inny wektor parametrów dla każdej osoby (heterogeniczność preferencji)
- ▶ Za mało informacji w danych by oszacować parametry dla każdego (podobnie jak z efektami stałymi, tylko gorzej)
- ▶ Zakładamy więc, że są to parametry losowe – każdy ma inne wartości parametrów, ale pochodzą one z założonego rozkładu
- ▶ Taki model nazywa się modelem parametrów losowych

Model parametrów losowych

- ▶ Warunkowe prawdopodobieństwo w k -tej sytuacji wyboru dane przez wzór na MNL

$$P(Y_{ijk} = 1 | \beta_i) = \frac{\exp(\mathbf{X}_{ijk} \beta_i)}{\sum_l \exp(\mathbf{X}_{ilk} \beta_i)}$$

- ▶ Ale parametry są nam nieznane, musimy uzyskać prawdopodobieństwo bezwarunkowe
 - ▶ Liczymy całkę (wielowymiarową)
 - ▶ Z powodu parametrów losowych, prawdopodobieństwa dla różnych sytuacji wyboru są zależne

$$P(Y_{ij_1 1}, Y_{ij_2 2}, \dots, Y_{ij_T T}) = \int \prod_t P(Y_{ijt} | \beta_i) f(\beta_i | \theta) d\beta_i$$

Model parametrów losowych

- ▶ łączne prawdopodobieństwo dane przez:

$$P(Y_{ij_1 1}, Y_{ij_2 2}, \dots, Y_{ij_T T}) = \int \prod_t P(Y_{ij_t t} | \beta_i) f(\beta_i | \theta) d\beta_i$$

- ▶ $f(\beta_i | \theta)$ to gęstość założonego rozkładu. Zależy ona od parametrów, które będziemy estymować
- ▶ Estymacja metodą Symulowanej Największej Wiarygodności
 - ▶ Całkę przybliżamy liczbami pseudo-, albo quasi-losowymi

$$L_i = P(Y_{ij_1 1}, Y_{ij_2 2}, \dots, Y_{ij_T T}) \approx \frac{1}{R} \sum_r \prod_t P(Y_{ij_t t} | \beta_i^r)$$

Zadanie 2. Sortowanie śmieci

1. Wczytaj zbiór danych me.recycling.dta
2. Dokonaj estymacji MNL
3. Dokonaj estymacji modelu z parametrami losowymi zakładając normalne rozkłady
 1. Oszacuj średnie WTP
 2. Czy wyniki będą takie same dla różnych ziaren?
4. Dokonaj estymacji modelu z parametrami losowymi zakładając rozkład log-normalny dla kosztu
 1. Ponownie oszacuj średnie WTP



Model klas ukrytych

- ▶ Alternatywą jest tzw. model klas ukrytych
 - ▶ Zamiast zakładać jakiś określony rozkład ciągły, zakładamy rozkład dyskretny
 - ▶ Parametry przyjmują np. dwie dowolne wartości z pewnymi prawdopodobieństwami
 - ▶ Trochę tak, jakbyśmy zakładali, że istnieją pewne segmenty rynku
- ▶ Funkcja wiarygodności dana przez

$$L_i = P(Y_{i_1 1}, Y_{i_2 2}, \dots, Y_{i_T T}) = \sum_c \pi_c \prod_t P(Y_{i_t t} | \boldsymbol{\beta}_c)$$

- ▶ π_c to prawdopodobieństwo przynależności dla klasy c

Zadanie 2. Sortowanie śmieci c.d.

5. Dokonaj estymacji modelu klas ukrytych

- ▶ Z dwiema klasami
- ▶ Z trzema klasami
- ▶ Porównaj wyniki
- ▶ Analogiczne możliwości dla innych modeli

- ▶ `fmm` – funkcja pozwalająca na estymację modeli z klasami ukrytymi dla różnych rozkładów, m.in. Poissona albo Ujemnego Dwumianowego
- ▶ `mepoisson` – model Poissona z parametrami losowymi