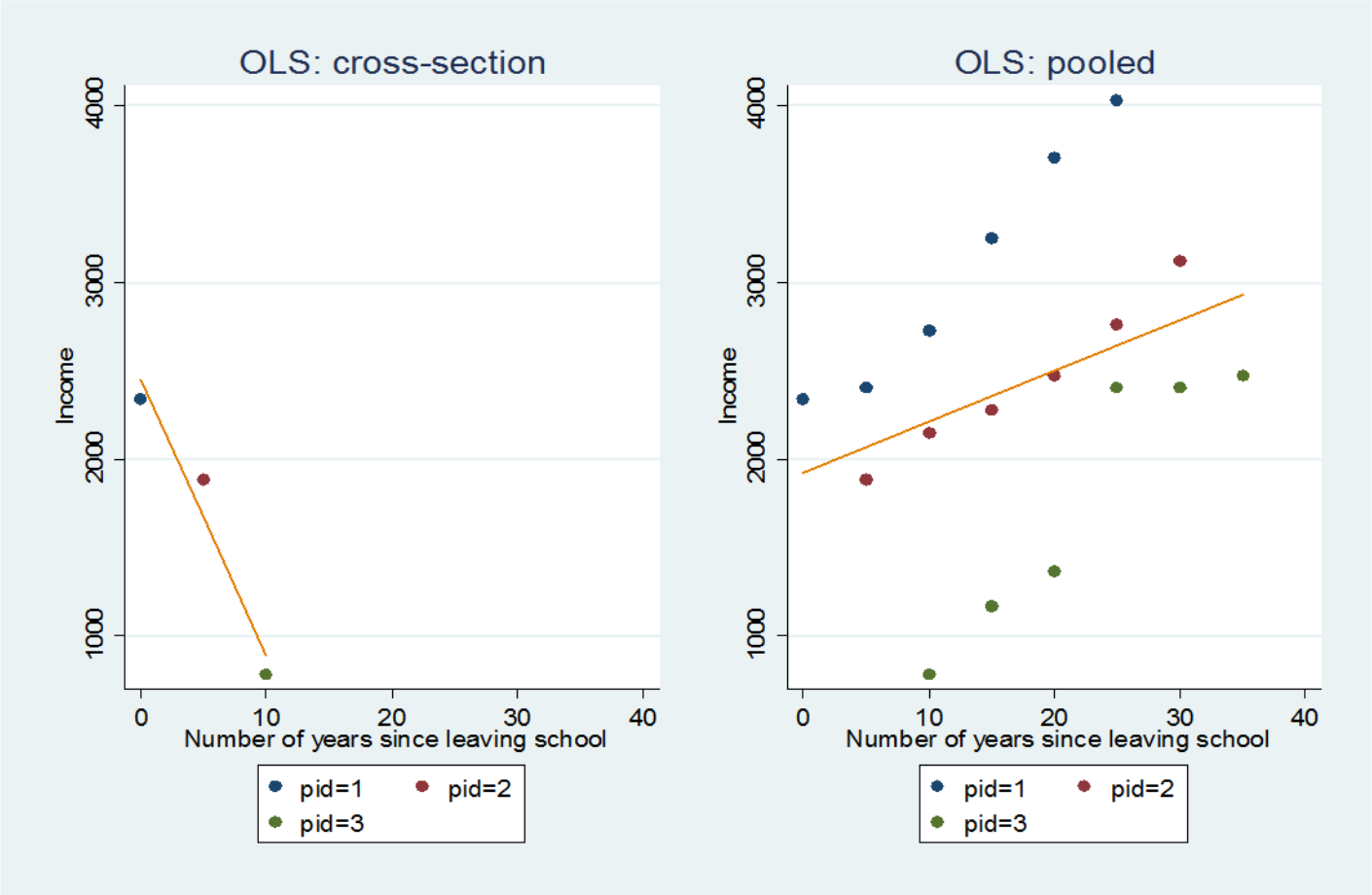
Analiza danych panelowych

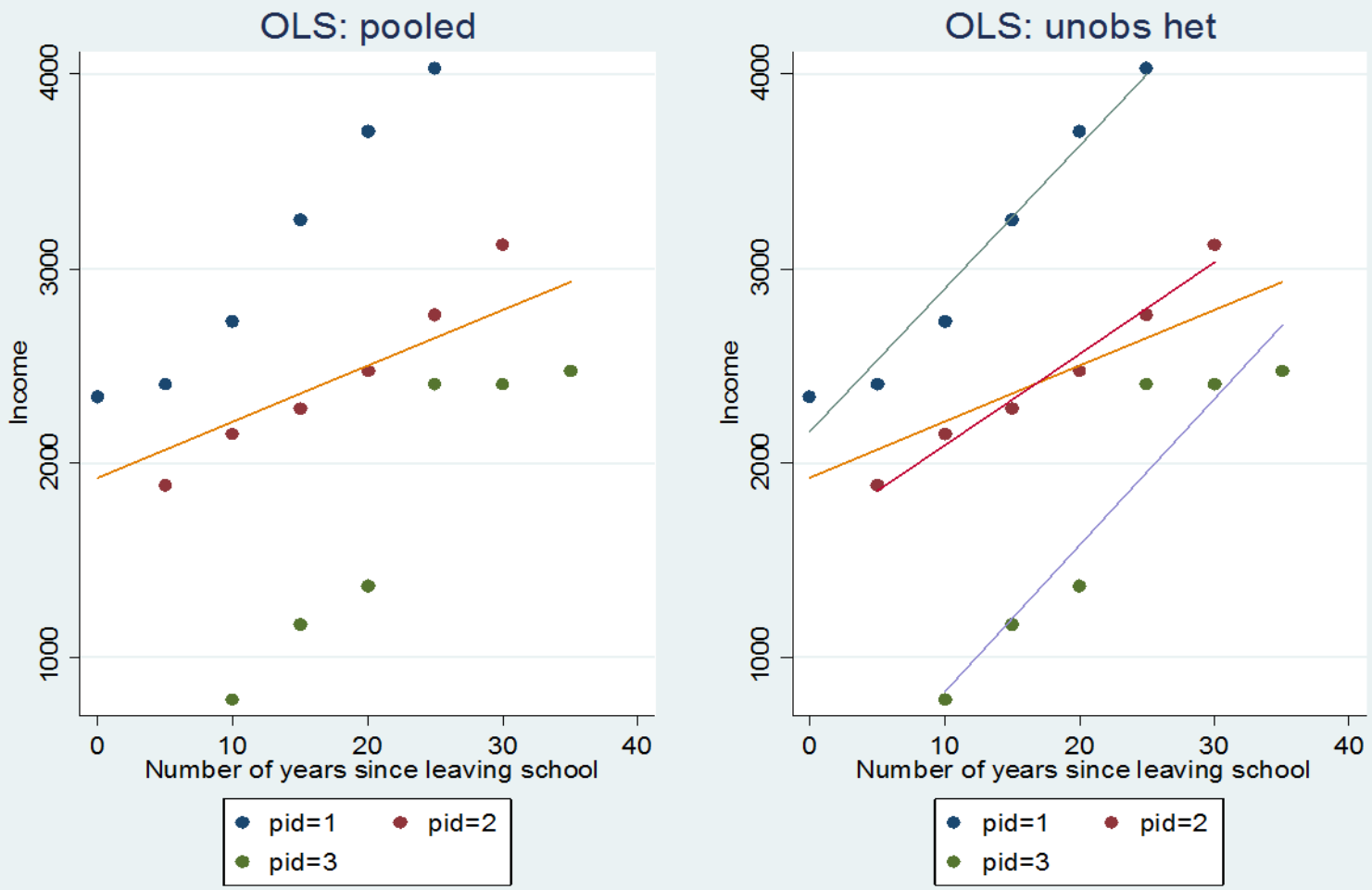
Stanisław Cichocki

Natalia Nehrebecka

Plan zajęć


OLS: cross-section
Income
1000
2000
3000
4000
0
10
20
30
40
Number of years since leaving school
pid=1
pid=2
pid=3
OLS: pooled
Income
1000
2000
3000
4000
0
10
20
30
40
Number of years since leaving school
pid=1
pid=2
pid=3

Plan zajęć


OLS: pooled
Income
1000
2000
3000
4000
0
10
20
30
40
Number of years since leaving school
pid=1
pid=2
pid=3
OLS: unobs het
Income
1000
2000
3000
4000
0
10
20
30
40
Number of years since leaving school
pid=1
pid=2
pid=3

Plan zajęć

1. **Wprowadzenie do danych panelowych**
   a) Charakterystyka danych panelowych
   b) Zalety i ograniczenia

2. **Modele ekonometryczne danych panelowych**
   a) Model efektów nieobserwowalnych
   b) Model efektów losowych
   c) Model efektów stałych
   d) Modele z dwukierunkowym komponentem błędu

3. **Testy**
   a) Test istotności efektów stałych i losowych
   b) Test Hausmana

Plan zajęć

1. Wprowadzenie do danych panelowych
   a) Charakterystyka danych panelowych
   b) Zalety i ograniczenia

2. Modele ekonometryczne danych panelowych
   a) Model efektów nieobserwowalnych
   b) Model efektów losowych
   c) Model efektów stałych
   d) Modele z dwukierunkowym komponentem błędu

3. Testy
   a) Test istotności efektów stałych i losowych
   b) Test Hausmana

Dane statystyczne

- Dane statystyczne dzieli się na trzy główne grupy:
- Szeregi czasowe
- Dane przekrojowe
- Dane panelowe
  - Dzielone są na dwie kategorie:
  - Dane „połączone” ($pooled\ data$)
  - Dane panelowe (właściwe)
    - Panel / longitudinal data

Próby przekrojowo-czasowe

- Dane „połączone” ($pooled\ data$)
- Dane z kilku okresów czasu dotyczące ***różnych*** prób przekrojowych
- Dane „połączone” to taki rodzaj danych, w którym nie rozróżniamy kolejnych jednostek badania.
- **Korzyści:** zwiększenie liczby stopni swobody z N na NT (N – liczba obserwacji, T – liczba okresów)

Próby przekrojowo-czasowe

obsno	*year*	*hprice*	*proptax*	*sqrft*	*bdrms*	*bthrms*
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Dane panelowe

- Dane panelowe właściwe
  - Dane panelowe mają cechy zarówno próby przekrojowej, jak i szeregu czasowego.
  - Zawierają one szereg obserwacji dla danej próby przekrojowej.
  - Każda obserwacja przypisana jest jednej konkretnie jednostce badania za pomocą zmiennej, która identyfikuje tę jednostkę.
  - Jednostki badania mogą być różnych rozmiarów: np. kraj, firma, rodzina, poszczególne osoby.
  - W danych panelowych występuje także zmienna wskazująca na okres, w którym dana obserwacja była zebrana (kwartał, rok, itd.).

Dane panelowe

A Two-Year Panel Data Set on City Crime Statistics

obsno	*city*	*year*	*murders*	*population*	*unem*	*police*
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Dane panelowe - zbilansowane

```
. xtset country year
       panel variable:  country (strongly balanced)
        time variable:  year, 1990 to 1999
                delta:  1 unit
```

Dane panelowe - zbilansowane

xtdes

```
 country:  1, 2, ..., 7                                        n =          7
    year:  1990, 1991, ..., 1999                               T =         10
           Delta(year) = 1 unit
           Span(year)  = 10 periods
           (country*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        10      10      10        10        10      10      10

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+------------
        7    100.00  100.00 |  1111111111
 ---------------------------+------------
        7    100.00         |  XXXXXXXXXX
```

Dane panelowe - niezbilansowane

. *xtset idcode year*

```
panel variable:  idcode (unbalanced)
time variable:  year, 68 to 88, but with gaps
          delta:  1 unit
```

Dane panelowe - niezbilansowane

xtdes

```
  idcode:  1, 2, ..., 5159                                    n =       4709
    year:  68, 69, ..., 88                                    T =         15
           Delta(year) = 1 unit
           Span(year)  = 21 periods
           (idcode*year uniquely identifies each observation)
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         5         9      13      15

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------------
      136      2.89    2.89 |  1....................
      114      2.42    5.31 |  .....................1
       89      1.89    7.20 |  ..................1.11
       87      1.85    9.05 |  ....................11
       86      1.83   10.87 |  111111.1.11.1.11.1.11
       61      1.30   12.17 |  ...............11.1.11
       56      1.19   13.36 |  11...................
       54      1.15   14.50 |  ................1.1.11
       54      1.15   15.65 |  .......1.11.1.11.1.11
     3972     84.35  100.00 | (other patterns)
 ---------------------------+-----------------------
     4709    100.00         |  XXXXXX.X.XX.X.XX.X.XX
```

Plan zajęć

1. Wprowadzenie do danych panelowych
   a) Charakterystyka danych panelowych
   b) Zalety i ograniczenia

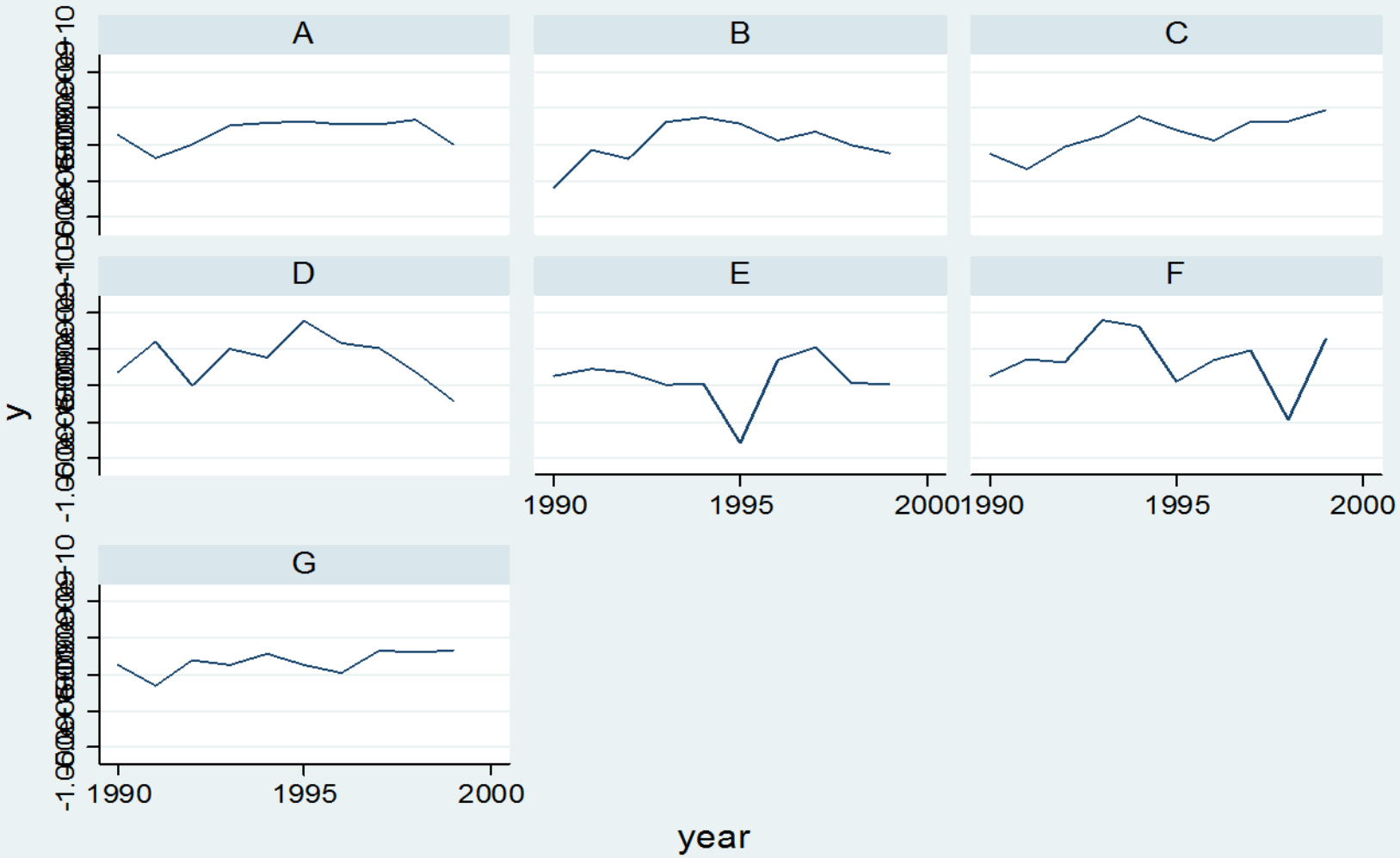
2. Modele ekonometryczne danych panelowych
   a) Model efektów nieobserwowalnych
   b) Model efektów losowych
   c) Model efektów stałych
   d) Modele z dwukierunkowym komponentem błędu

3. Testy
   a) Test istotności efektów stałych i losowych
   b) Test Hausmana

Dane panelowe - zbilansowane

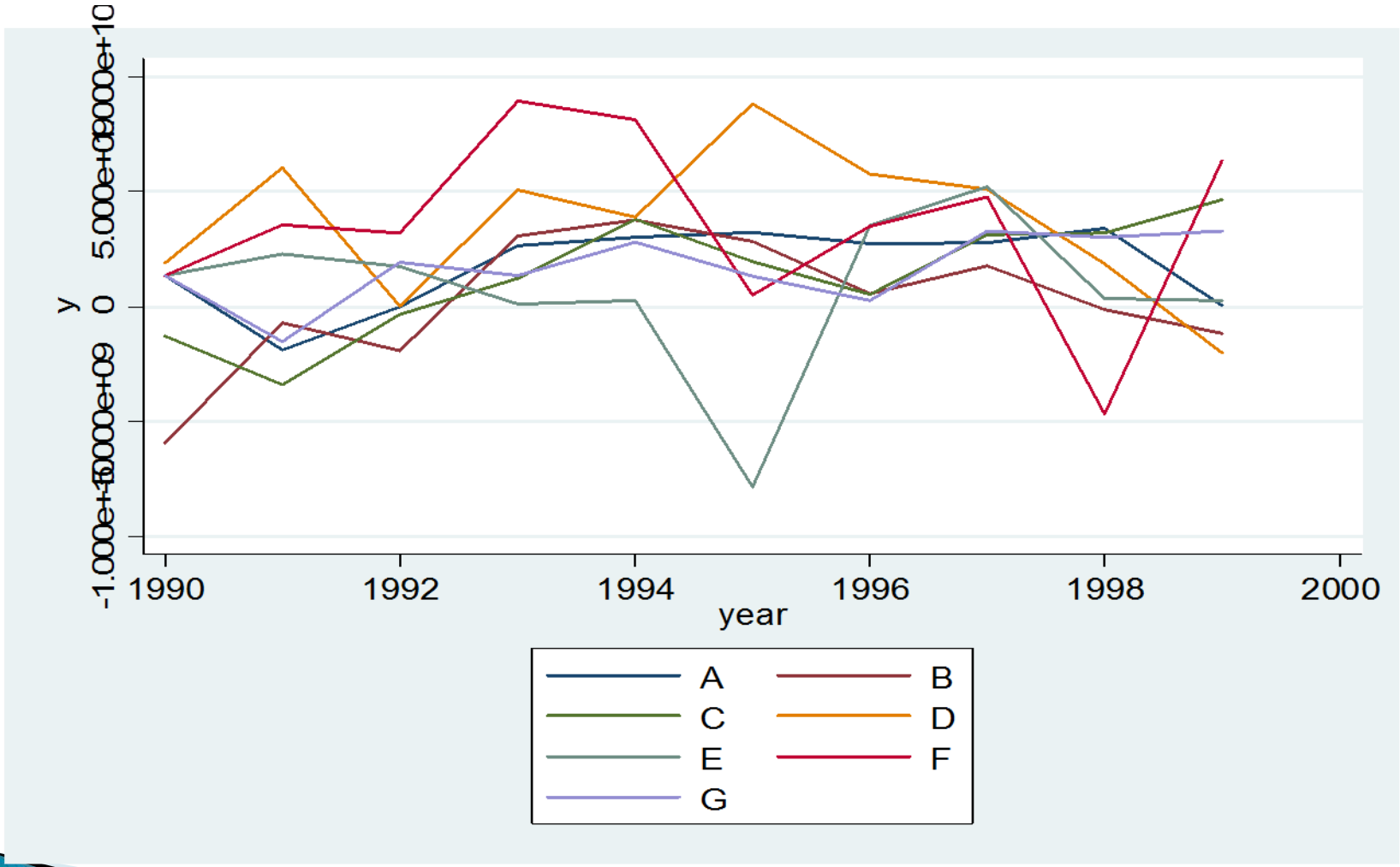
```
. xtset country year
       panel variable:  country (strongly balanced)
        time variable:  year, 1990 to 1999
                delta:  1 unit
```

Dane panelowe



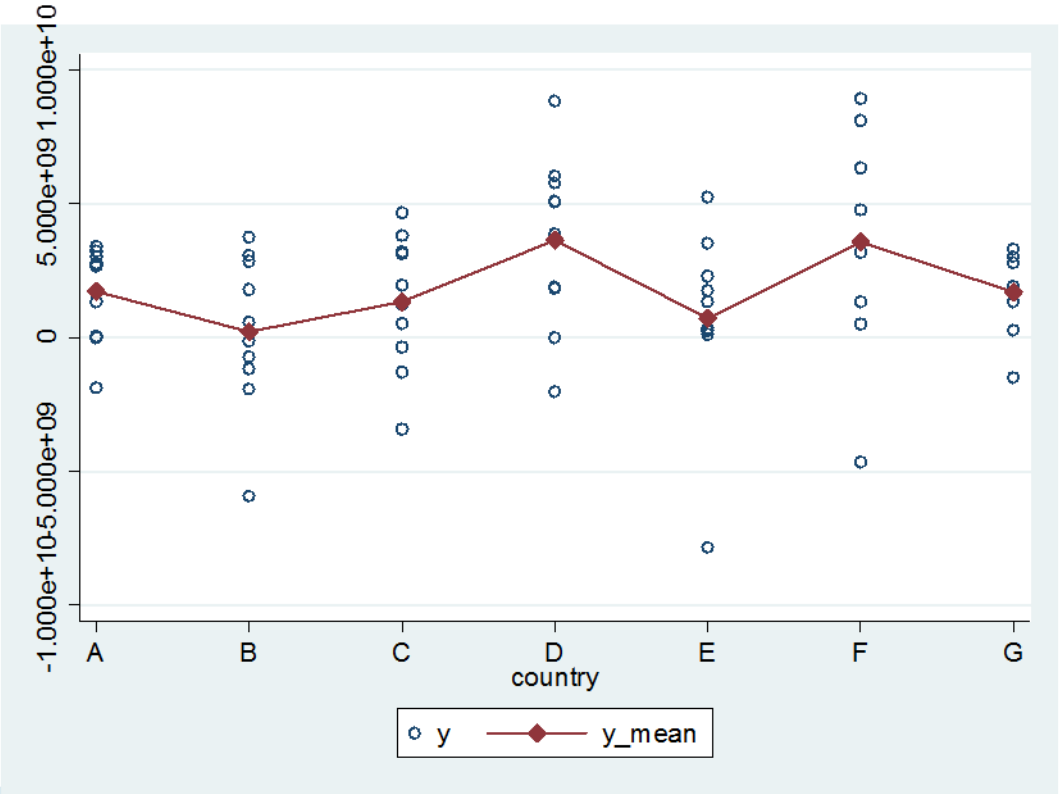
Graphs by country

Dane panelowe



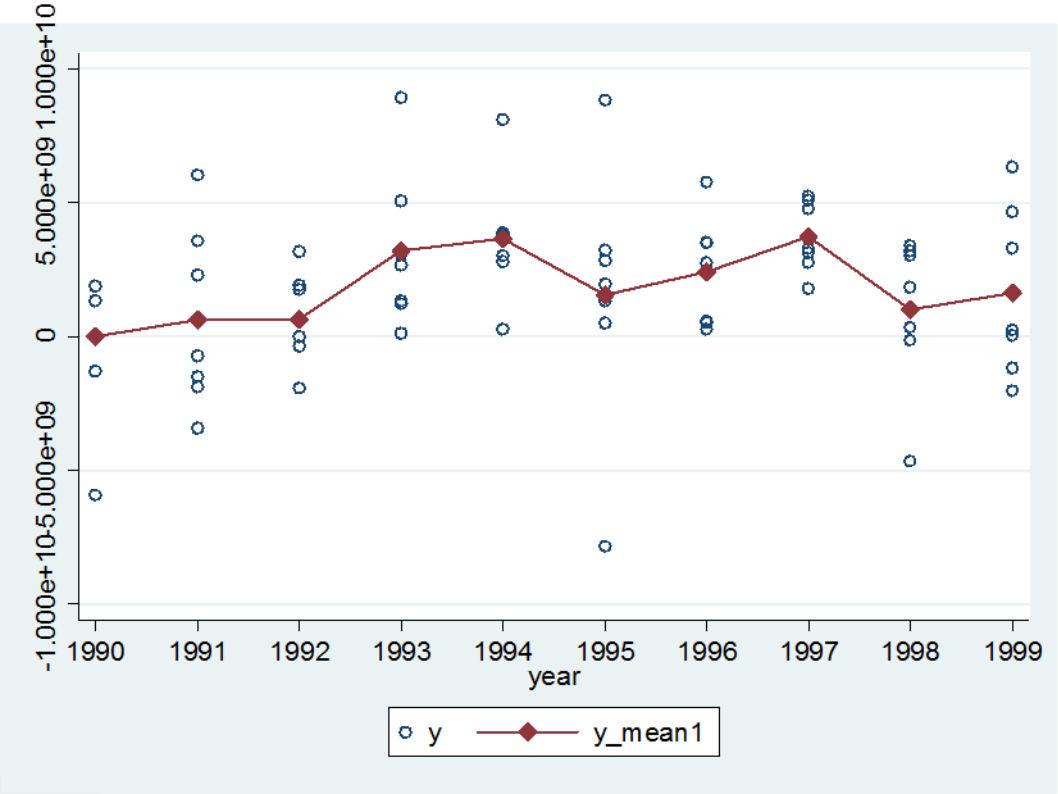
Dane panelowe – zalety

- Kontrolowanie heterogeniczności jednostek badania
  - *Heterogeneity across countries*



Dane panelowe – zalety

- Kontrolowanie heterogeniczności jednostek badania
  - *Heterogeneity across years*



Dane panelowe – zalety

- Kontrolowanie heterogeniczności jednostek badania
- Dane panelowe mają większą moc informacyjną, większą różnorodność, **mniejszą współliniowość między zmiennymi**, więcej stopni swobody oraz większą efektywność.
- Dane panelowe są bardziej odpowiednie do **badania dynamiki dostosowań**.
- Dane panelowe bardziej nadają się do identyfikowania i mierzenia efektów, które w oczywisty sposób nie są możliwe do analizy za pomocą prób przekrojowych i szeregów czasowych.

Dane panelowe – ograniczenia

- Projekt kwestionariusza oraz proces zbierania danych
  - Niewystarczająca liczba jednostek, którymi badacz jest zainteresowany.
  - Brak odpowiedzi – brak współpracy respondenta lub błędy ankietera.
  - Częstość przeprowadzania wywiadów.

- Błędy pomiarowe

- Problemy związane z selekcją jednostek
  - **Brak odpowiedzi** – związany z odmową przystąpienia do badania, nieobecnością respondenta w domu, itd.
  - **„Wycieranie" panelu ($attrition$)** – jest to trwałe „wypadanie" jednostek objętych badaniem z takich powodów jak: śmierć, przeprowadzka, dłuższy wyjazd (np. na studia).

Dane panelowe – ograniczenia

Freq.	Percent	Cum.	Pattern
13674	83.86	83.86	11..11
952	5.84	89.70	.1..11
738	4.53	94.22	11
346	2.12	96.34	11....
232	1.42	97.77	1
142	0.87	98.64	11..1.
122	0.75	99.39	1.....
38	0.23	99.62	.1....
31	0.19	99.81	....1.
15	0.09	99.90	.1..1.
7	0.04	99.94	11...1
5	0.03	99.98	1...11
4	0.02	100.00	.1...1
16306	100.00		XX..XX

$$2.12 + 0.87 + 0.75 = 3.74\%$$

Plan zajęć

1. Wprowadzenie do danych panelowych
   a) Charakterystyka danych panelowych
   b) Zalety i ograniczenia

2. **Modele ekonometryczne danych panelowych**
   a) Model efektów nieobserwowalnych
   b) Model efektów losowych
   c) Model efektów stałych
   d) Modele z dwukierunkowym komponentem błędu

3. Testy
   a) Test istotności efektów stałych i losowych
   b) Test Hausmana

Ekonometryczne modele danych panelowych – liniowy model efektów nieobserwowalnych

- Liniowy model efektów nieobserwowalnych dla danych panelowych składających się z N jednostek i T okresów możemy zapisać następująco:

$$y_{it} = \mathbf{x_{it}}\boldsymbol{\beta} + \underbrace{u_i + \varepsilon_{it}}_{\boldsymbol{v_{it}}}$$

i	–	indeks obserwacji,
t	–	indeks czasu,
y_{it}	–	zmienna objaśniana dla i–tej jednostki w czasie t,
x_{it}	–	wektor obserwacji zmiennych objaśniających dla i–tej jednostki w czasie t,
β	–	wektor parametrów,
u_i	–	efekt indywidualny dla i–tej jednostki,
ε_{it}	–	błąd czysto losowy dla i–tej jednostki w czasie t,
v_{it}	–	łączny błąd losowy dla i–tej jednostki w czasie t

- Poznane przez nas modele będą różnić się koncepcją traktowania efektu indywidualnego

Ekonometryczne modele danych panelowych – liniowy model efektów nieobserwowalnych

- Model z jednokierunkowym komponentem błędu (***one way error component model***)

$$v_{it} = u_i + \varepsilon_{it}$$

u_i	–	efekt indywidualny dla i–tej jednostki,
ε_{it}	–	błąd czysto losowy dla i–tej jednostki w czasie t
v_{it}	–	łączny błąd losowy dla i–tej jednostki w czasie t

- Model z dwukierunkowym komponentem błędu (***two way error component model***)

$$v_{it} = u_i + \lambda_t + \varepsilon_{it}$$

λ_t	–	efekt czasowy

Plan zajęć

1. Wprowadzenie do danych panelowych
   a) Charakterystyka danych panelowych
   b) Zalety i ograniczenia

2. **Modele ekonometryczne danych panelowych**
   a) Model efektów nieobserwowalnych
   b) Model efektów losowych
   c) Model efektów stałych
   d) Modele z dwukierunkowym komponentem błędu

3. Testy
   a) Test istotności efektów stałych i losowych
   b) Test Hausmana

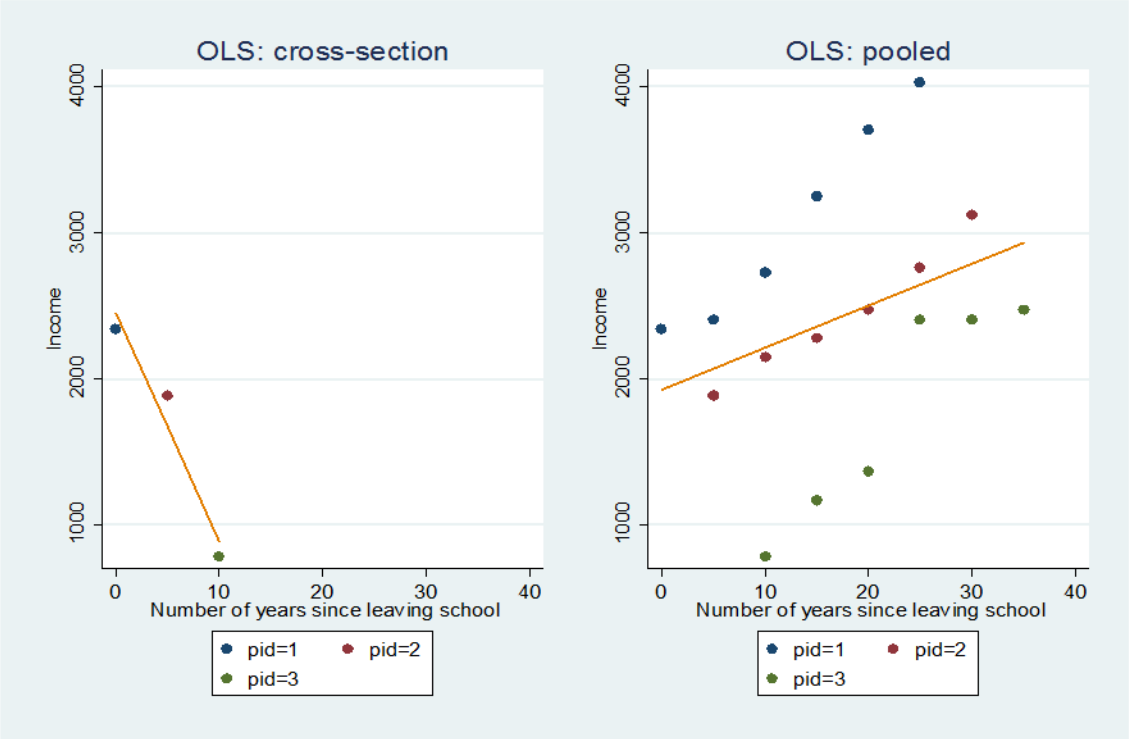
Ekonometryczne modele danych panelowych – pooled regression

- Regresja typu *pooled* ($Pooled\ Ordinary\ Least\ Square$), to regresja na danych, które nie rozróżniają jednostek badania i okresów czasu.
- Zapominamy, iż dane mają wymiar przekrojowo-czasowy i estymujemy wszystkie obserwacje łącznie za pomocą metody najmniejszych kwadratów.
- Regresja ta nie wykorzystuje struktury danych panelowych – **nie bierze pod uwagę heterogeniczności jednostek badania**.
- Wykorzystując metodę *pooled regresion* zakłada się, iż wszystkie obserwacje pochodzą z jednego okresu czasu, w wyniku czego w równaniu panelowym pojawia się łączny błąd losowy υ_{it} (odpowiadający sumie ε_{it} oraz u_i)

$$y_{it} = \boldsymbol{x_{it}\beta} + \upsilon_{it}, \qquad \upsilon_{it} = u_i + \varepsilon_{it}$$

Ekonometryczne modele danych panelowych – pooled regression

pid	wave	y	x1
1	1	2340	0
1	2	2405	5
1	3	2730	10
1	4	3250	15
1	5	3705	20
1	6	4030	25
2	1	1885	5
2	2	2145	10
2	3	2275	15
2	4	2470	20
2	5	2762	25
2	6	3120	30
3	1	780	10
3	2	1170	15
3	3	1365	20
3	4	2405	25
3	5	2405	30
3	6	2470	35



$OLS_{t\ =\ 1}$: $y = 2448 - 156 * x1$ $\quad OLS_{pooled}$: $y = 1925 + 29 * x1$

Ekonometryczne modele danych panelowych – pooled regression

- Aby otrzymany estymator był estymatorem **zgodnym**, gdy:

$$E(v_{it}) = 0, Cov(v_{it}, x_{it}) = 0$$

  - zakładamy, że błąd czysto losowy ε_{it} ma wartość oczekiwaną równą zero i jest nieskorelowany ze zmiennymi objaśniającymi:

$$E(\varepsilon_{it}) = 0, Cov(\varepsilon_{it}, x_{it}) = 0$$

  - musi być spełniony warunek braku korelacji pomiędzy efektami indywidualnymi a zmiennymi objaśniającymi:

$$E(u_i) = 0, \boldsymbol{Cov(u_i, x_{it}) = 0}$$

Ekonometryczne modele danych panelowych – pooled regression

- Estymator MNK jest estymatorem **nieobciążonym**, jeśli

$$E(\varepsilon_{it}|\boldsymbol{X}) = 0$$

$$E(u_i|\boldsymbol{X}) = 0$$

Ekonometryczne modele danych panelowych – pooled regression

- Otrzymany estymator MNK (*przy powyższych założeniach*) będzie estymatorem zgodnym, lecz **nieefektywnym.**
  - Na podstawie twierdzenia Gaussa-Markowa estymator MNK jest efektywny, jeśli błąd losowy jest homoskedastyczny i nieskorelowany. Sformułowano założenia o tym że:

$$Var(\boldsymbol{u} \mid \boldsymbol{X}) = {\sigma_u}^2 \boldsymbol{I}$$

$$Var(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = {\sigma_\varepsilon}^2 \boldsymbol{I}$$

$$Cov(\boldsymbol{u}, \boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0$$

- $\boldsymbol{\varepsilon} = [\varepsilon_{11} \quad \dots \quad \varepsilon_{1T} \quad \dots \quad \varepsilon_{N1} \quad \dots \quad \varepsilon_{NT}]; \ \mathbf{u} = [u_1 \quad \dots \quad u_N]$
  - **Czy założenie te implikują homoskedastyczność i brak autokorelacji łącznego błędu losowego v_{it}?**

Ekonometryczne modele danych panelowych – pooled regression

- Przy tych założeniach wariancja błędu łącznego wynosi:
  - Dla uproszczenie zapisu pomijamy warunkowanie względem $\boldsymbol{X}$

$$Var(v_{it}) = Var(u_i + \varepsilon_{it}) = Var(u_i) + Var(\varepsilon_{it}) + 2Cov(u_i, \varepsilon_{it})$$
$$= {\sigma_u}^2 + {\sigma_\varepsilon}^2$$

- Dla $t \neq s$, kowariancja między v_{it} i v_{is} jest równa

$$\boldsymbol{Cov(v_{\mathrm{it}}\,,\, v_{\mathrm{is}}) = {\sigma_u}^2}$$

- Dla $i \neq j$, kowariancja między v_{it} i v_{jt} jest równa

$$Cov(v_{it}\,,\, v_{\mathrm{jt}}) = 0$$

- W rezultacie macierz wariancji-kowariancji błędu losowego nie jest sferyczna!

Ekonometryczne modele danych panelowych – pooled regression

- Przy spełnieniu założeń $E(u_i) = 0, Cov(u_i, x_{it}) = 0$ oraz w związku z występowaniem korelacji pomiędzy łącznymi błędami losowymi można stwierdzić, że oszacowany **estymator parametrów modelu *pooled regresion* będzie zgodny, lecz nieefektywny**.

- Powinno się w tym wypadku użyć **estymatora warstwowego**.
  - W związku z tym wnioskowanie statystyczne przeprowadza się na podstawie estymacji, w której wykorzystano **macierz odporną wariancji-kowariancji**.

Pytania teoretyczne

1. Czym różni się panel od próby przekrojowo-czasowej?

2. Zapisać liniowy model efektów nieobserwowalnych i zinterpretować jego elementy.

Dziękuję za uwagę