

Analiza danych panelowych cz II

Stanisław Cichocki

Natalia Nehrebecka

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

- ▶ NLSY data on young women aged 14–26 years in 1968
- ▶ Opis zmiennych:
 - *idcode* - identyfikator jednostki
 - *year* - rok
 - *age* - wiek w latach
 - *race* - rasa: 1=biały, 2=czarny, 3=inny
 - *grade* - wykształcenie mierzone liczbą lat nauki
 - *not_smsa* - 1 jeśli osoba nie mieszka w obszarze miejskim (SMSA - *Standard Metropolitan Statistical Area*)
 - *tth_exp* - doświadczenie zawodowe mierzone liczbą lat nauki
 - *tenure* - liczba przepracowanych lat w obecnym miejscu zatrudnienia
 - *ln_wage* - logarytm(płac/deflator)

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

xtides

```

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	5.311
89	1.89	7.201.11
87	1.85	9.0511
86	1.83	10.87	111111.1.11.1.11.1.11
61	1.30	12.1711.1.11
56	1.19	13.36	11.....
54	1.15	14.501.1.11
54	1.15	15.651.11.1.11.1.11
3972	84.35	100.00	(other patterns)
4709	100.00		XXXXXX.X.XX.X.XX.X.XX

Ekonometryczne modele danych panelowych – pooled regression

Source	SS	df	MS	Number of obs	=	28089
Model	2273.11093	7	324.730133	F(7, 28081)	=	2205.19
Residual	4135.1272	28081	.14725712	Prob > F	=	0.0000
				R-squared	=	0.3547
				Adj R-squared	=	0.3546
Total	6408.23813	28088	.228148609	Root MSE	=	.38374

ln_wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0042929	.0005277	-8.14	0.000	-.0053273	-.0032586
_Irace_2	-.0947803	.0051932	-18.25	0.000	-.1049593	-.0846013
_Irace_3	.0525272	.0222947	2.36	0.018	.0088286	.0962258
grade	.0666069	.0010359	64.30	0.000	.0645764	.0686374
not_smsa	-.1915868	.0051516	-37.19	0.000	-.2016842	-.1814895
t1l_exp	.0285226	.0009286	30.72	0.000	.0267026	.0303426
tenure	.0205235	.0008356	24.56	0.000	.0188857	.0221613
_cons	.8052863	.0176248	45.69	0.000	.7707407	.8398319

Ekonometryczne modele danych panelowych – pooled regression

► First-order autocorrelation in a variable

sort idcode year

```
. correlate ln_wage L.ln_wage
      |                               L.
      |   ln_wage   ln_wage
-----+-----
      |
ln_wage |
  --.   |   1.0000
  L1.   |   0.7557   1.0000
```

► Autocorrelations of residual

```
Autocorrelation at lag 1 = 0.635
Autocorrelation at lag 2 = 0.589
Autocorrelation at lag 3 = 0.529
Autocorrelation at lag 4 = 0.463
Autocorrelation at lag 5 = 0.442
Autocorrelation at lag 6 = 0.374
```

Ekonometryczne modele danych panelowych – pooled regression

- ▶ **First-order autocorrelation differs in different year pairs**

Autocorrelation at lag 1 in year 2 = 0.558

Autocorrelation at lag 1 in year 3 = 0.679

Autocorrelation at lag 1 in year 4 = 0.582

Autocorrelation at lag 1 in year 5 = 0.586

Autocorrelation at lag 1 in year 6 = 0.715

Ekonometryczne modele danych panelowych – pooled regression

► Pooled OLS with cluster-robust standard errors

```
xi: reg ln_wage age i.race grade not_smsa ttl_exp tenure, cluster(idcode)
```

```
i.race          _Irace_1-3          (naturally coded; _Irace_1 omitted)
```

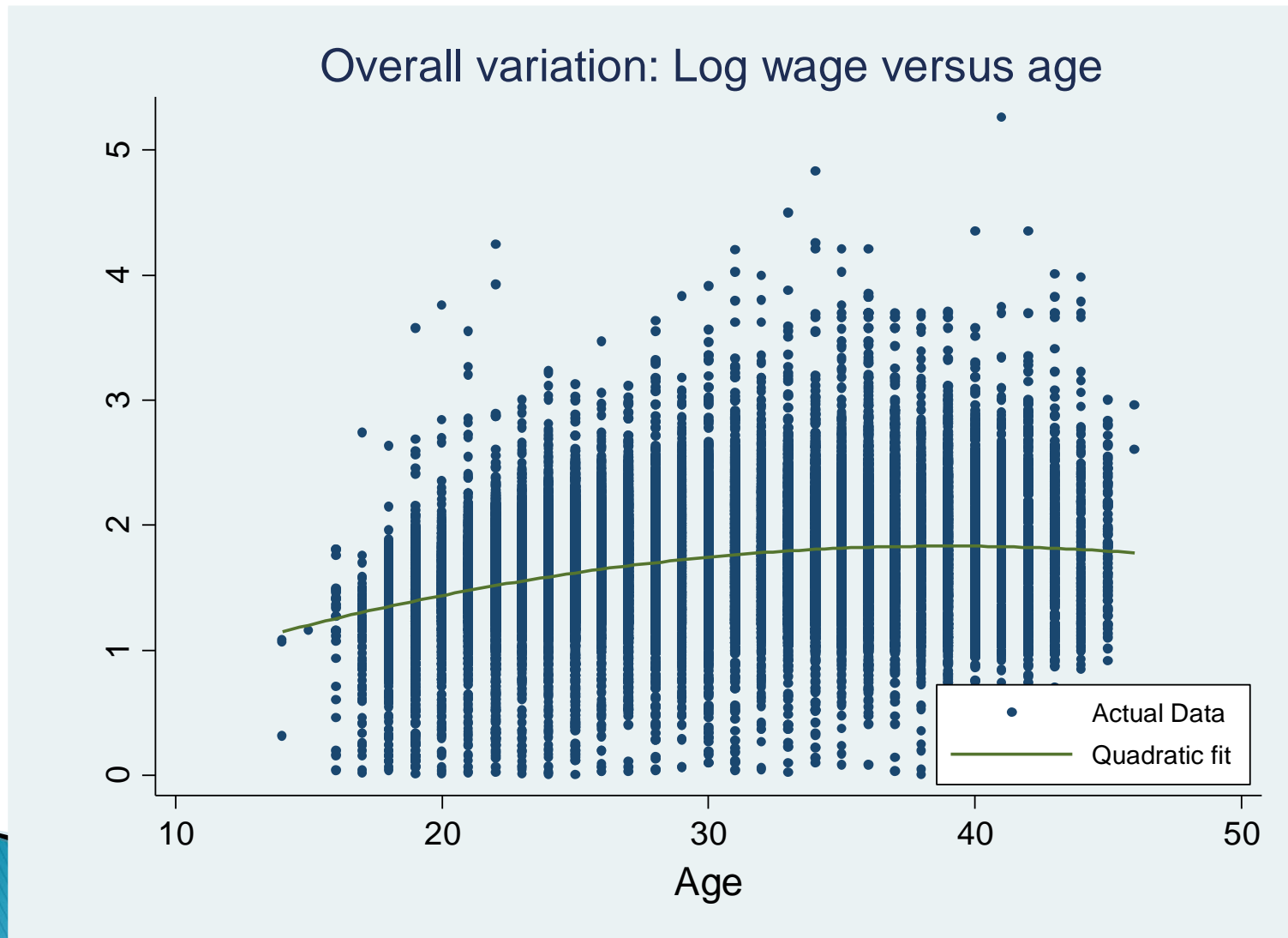
Linear regression

Number of obs = 28089
F(7, 4694) = 696.39
Prob > F = 0.0000
R-squared = 0.3547
Root MSE = .38374

(Std. Err. adjusted for 4695 clusters in idcode)

		Robust				
ln_wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0042929	.0009076	-4.73	0.000	-.0060722	-.0025137
_Irace_2	-.0947803	.010053	-9.43	0.000	-.1144889	-.0750717
_Irace_3	.0525272	.0474282	1.11	0.268	-.0404543	.1455087
grade	.0666069	.0020774	32.06	0.000	.0625342	.0706797
not_smsa	-.1915868	.009542	-20.08	0.000	-.2102936	-.1728801
ttl_exp	.0285226	.0017769	16.05	0.000	.0250391	.0320061
tenure	.0205235	.0015556	13.19	0.000	.0174738	.0235732
_cons	.8052863	.0333542	24.14	0.000	.7398963	.8706762

Ekonometryczne modele danych panelowych – pooled regression



Ekonometryczne modele danych panelowych – pooled regression

► Pooled OLS with cluster-robust standard errors

reg ln_wage age age_2 black grade not_smsa ttl_exp tenure, cluster(idcode)

Linear regression

Number of obs = 28089
F(7, 4694) = 725.78
Prob > F = 0.0000
R-squared = 0.3602
Root MSE = .38211

(Std. Err. adjusted for 4695 clusters in idcode)

ln_wage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0431842	.0040712	10.61	0.000	.0352028	.0511657
age_2	-.0007933	.0000697	-11.38	0.000	-.00093	-.0006567
black	-.0997647	.0100641	-9.91	0.000	-.1194951	-.0800343
grade	.0648427	.0020844	31.11	0.000	.0607563	.068929
not_smsa	-.1921946	.0095236	-20.18	0.000	-.2108652	-.173524
ttl_exp	.0288935	.0017859	16.18	0.000	.0253923	.0323947
tenure	.0203773	.0015568	13.09	0.000	.0173253	.0234293
_cons	.1535139	.0609553	2.52	0.012	.0340128	.273015

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 – random effects model

- ▶ Zastosowanie modelu z efektami losowymi - *RE* (ang. *random effects model*) implikuje **założenie o braku korelacji pomiędzy efektami indywidualnymi u_i a zmiennymi objaśniającymi x_{it} .**

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

- ▶ W celu otrzymania zgodnego i efektywnego estymatora efektów losowych nakładane są dodatkowe warunki dotyczące składnika losowego oraz efektów indywidualnych.

Ekonometryczne modele danych panelowych – random effects model

- ▶ Model efektów losowych ma następujące założenia:
- ▶ **Zgodność** estymatora wymaga, by dla poszczególnych jednostek zmienne objaśniające nie były skorelowane z charakterystykami indywidualnymi i błędami czysto losowymi pochodzącymi z różnych okresów czasu.

- Brak jednoczesności (dokładna egzogeniczność)

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

- Niezależność:

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

Ekonometryczne modele danych panelowych – random effects model

- ▶ założenia dodatkowe – potrzebne do udowodnienia **efektywności** estymatora parametrów tego modelu
 - brak korelacji pomiędzy błędem czysto losowym i efektem indywidualnym:

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

- obowiązuje założenie o homoskedastyczności i braku autokorelacji:

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

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

Ekonometryczne modele danych panelowych – random effects model

- ▶ Zastanówmy się nad błędem losowym tak wyspecyfikowanego modelu.

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

- ▶ Czy błąd spełnia założenia Klasycznego Modelu Regresji Liniowej (KMRL), czyli czy jest szansa na skorzystanie z metody najmniejszych kwadratów.
 - W szczególności sprawdzimy założenie dotyczące sferyczności tego błędu.
 - Sferyczność oznacza brak autokorelacji i heteroskedastyczności

Ekonometryczne modele danych panelowych – random effects model

- ▶ Przy tych założeniach wariancja błędu łącznego wynosi:
 - Dla uproszczenia zapisu pomijamy warunkowanie względem X

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

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

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

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

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

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

Ekonometryczne modele danych panelowych – random effects model

- ▶ W rezultacie wnioskujemy, iż macierz wariancji kowariancji łącznego błędu losowego dla i -tej jednostki ma postać:

$$\Omega_i = Var(\mathbf{v}_i) = E(\mathbf{v}_i \mathbf{v}_i') = \begin{bmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \cdots & \sigma_u^2 \\ \vdots & \ddots & \vdots \\ \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

- ▶ Na mocy założeń zachodzi:

$$\Omega_1 = \Omega_2 = \cdots \Omega_N = \Omega$$

Ekonometryczne modele danych panelowych – random effects model

- ▶ Korzystając dalej z faktu, iż

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

- ▶ Mamy

$$\bar{\Omega} = Var(\mathbf{v}) = \begin{bmatrix} Var(\mathbf{v}_1) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & Var(\mathbf{v}_N) \end{bmatrix} = \begin{bmatrix} \Omega & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Omega \end{bmatrix}$$

- ▶ Wniosek: macierz wariancji-kowariancji błędu losowego nie jest diagonalna.
- ▶ Powyższa macierz będzie miała postać blokowo-diagonalną.

Ekonometryczne modele danych panelowych – random effects model

- ▶ Przeprowadzone wnioskowanie wskazuje, że otrzymana macierz wariancji i kowariancji składnika losowego ma postać blokowo - diagonalną, co pociąga za sobą użycie **Uogólnionej Metody Najmniejszych Kwadratów** (*Generalized Least Squares - GLS*).
- ▶ Z ogólnego wzoru na estymator *UMNK* wnioskujemy, iż estymatorem efektywnym w analizowanym przez nas przypadku jest estymator postaci:

$$\hat{\beta} = (X'\bar{\Omega}^{-1}X)^{-1}X'\bar{\Omega}^{-1}y$$

- ▶ Dzięki czemu korygowane są nieprawidłowości związane z niesferycznością macierzy $\bar{\Omega}$.

Ekonometryczne modele danych panelowych – random effects model

- Korzystając ze specyfiki macierzy $\bar{\Omega}$, łatwo pokazać, iż:

$$X' \bar{\Omega}^{-1} X = \sum_{i=1}^N X'_i \Omega^{-1} X_i$$

oraz

$$X' \bar{\Omega}^{-1} y = \sum_{i=1}^N X'_i \Omega^{-1} y_i$$

- Wtedy estymator *UMNK* przybiera atrakcyjniejszą obliczeniowo postać:

$$\hat{\beta} = \left(\sum_{i=1}^N X'_i \Omega^{-1} X_i \right)^{-1} \sum_{i=1}^N X'_i \Omega^{-1} y_i$$

Ekonometryczne modele danych panelowych – random effects model

- ▶ Kolejnym standardowym problem do rozwiązania, pojawiającym się przy zastosowaniu *UMNK* jest fakt nieznaności macierzy Ω .
- ▶ Należy oszacować macierz Ω , co zmienia metodę, z której będziemy korzystali z *UMNK* na *SUMNK* (*Stosowalną UMNK – Feasible – FGLS*).
- ▶ Oszacowanie macierzy Ω polega na oszacowaniu jej elementów.
 - Szacujemy *pooled regression*.
 - Wiemy, iż $\text{Var}(v_{it}) = \sigma_v^2 = \sigma_u^2 + \sigma_\varepsilon^2$
 - Jeśli spełnione są dwa pierwsze założenia modelu z losowymi efektami, to otrzymamy zgodny estymator postaci (bo oszacowania parametrów z *pooled regression* są zgodne):
$$\hat{\sigma}_v^2 = \frac{1}{NT - K} \sum_{i=1}^N \sum_{t=1}^T \hat{v}_{it}^2$$
 - gdzie \hat{v}_{it} są resztami z *pooled regression*.

Ekonometryczne modele danych panelowych – random effects model

- ▶ Wykorzystując fakt, iż $Cov(v_{it}, v_{is}) = \sigma_u^2$, wyznaczamy:

$$\hat{\sigma}_u^2 = \frac{1}{\frac{NT(T-1)}{2} - K} \sum_{i=1}^N \sum_{t=1}^{T-1} \sum_{s=t+1}^T \hat{v}_{it} \hat{v}_{is}$$

- ▶ A pamiętając, iż $\sigma_\varepsilon^2 = \sigma_v^2 - \sigma_u^2$, mamy w końcu wyznaczone oszacowania wszystkich elementów macierzy $\mathbf{\Omega}$, a tym samym oszacowaliśmy macierz $\hat{\mathbf{\Omega}}$.

Ekonometryczne modele danych panelowych – random effects model

- ▶ Oszacowanie macierz wariancji kowariancji łącznego błędu losowego dla i -tej jednostki ma postać:

$$\hat{\Omega}_i = \hat{\Omega} = \begin{bmatrix} \hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 & \hat{\sigma}_u^2 & \dots & \hat{\sigma}_u^2 \\ \hat{\sigma}_u^2 & \hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 & \dots & \hat{\sigma}_u^2 \\ \vdots & \vdots & \ddots & \hat{\sigma}_u^2 \\ \hat{\sigma}_u^2 & \dots & \hat{\sigma}_u^2 & \hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 \end{bmatrix}$$

Ekonometryczne modele danych panelowych – random effects model

- ▶ A estymator *SUMNK* wyznaczany jest ze wzoru:

$$\widehat{\beta}_{RE} = \left(\sum_{i=1}^N X'_i \hat{\Omega}^{-1} X_i \right)^{-1} \sum_{i=1}^N X'_i \hat{\Omega}^{-1} y_i$$

- Gdzie „RE” oznacza, iż jest to estymator modelu z losowymi efektami (i jest efektywniejszy od b_{MKN}).

Ekonometryczne modele danych panelowych – random effects model

xtreg ln_wage age age_2 black grade not_smsa ttl_exp tenure, re

Random-effects GLS regression

Group variable: idcode

R-sq: within = 0.1595
between = 0.4676
overall = 0.3567

Number of obs = 28089

Number of groups = 4695

Obs per group: min = 1
avg = 6.0
max = 15

corr(u_i, X) = 0 (assumed)

Wald chi2(7) = 8629.06

Prob > chi2 = 0.0000

ln_wage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0452213	.0026291	17.20	0.000	.0400685	.0503742
age_2	-.0008463	.0000437	-19.36	0.000	-.0009319	-.0007606
black	-.0783429	.009869	-7.94	0.000	-.0976857	-.059
grade	.0664253	.0017922	37.06	0.000	.0629126	.069938
not_smsa	-.1430699	.0071603	-19.98	0.000	-.1571039	-.1290359
ttl_exp	.0329041	.0011253	29.24	0.000	.0306986	.0351096
tenure	.0131911	.0008403	15.70	0.000	.011544	.0148381
_cons	.0923273	.0440066	2.10	0.036	.006076	.1785786

sigma_u | .26027187

sigma_e | .29278824

rho | .1140898

(fraction of variance due to u_i)

Pytania teoretyczne

1. Przy jakich założeniach estymator efektów losowych jest zgodny?
2. Z jakiego powodu estymując model za pomocą estymatora efektów losowych używamy Stosowalną Uogólnioną Metodę Najmniejszych Kwadratów?

Dziękuję za uwagę