

# 패널자료 회귀분석: 구조 및 분석 전략

황선재

충남대학교 사회학과 부교수  
sunjaeh@cnu.ac.kr

# 목차

## ▶ 패널자료 소개

- ▶ 패널자료 정의 및 예시
- ▶ 패널자료용 분석 기법이 필요한 이유
- ▶ 패널자료 분석의 장단점

## ▶ 패널자료 분석 전략

- ▶ 횡단자료 vs. 패널자료 회귀분석
- ▶ 패널자료 회귀분석 기법 종류
- ▶ 패널자료 회귀분석 기본 전략

## ▶ 패널자료 분석 기법

- ▶ 고정효과 모형(Fixed Effects Model: FE)
- ▶ 임의효과 모형(Random Effects Model: RE)

# 참고문헌 및 읽을거리

## ▶ 교재

- ▶ 한치록. 2021. 패널데이터강의(제3판). 박영사.
- ▶ 민인식·최필선. 2019. STATA 패널데이터분석(제2판). 지필미디어.

## ▶ 논문/보고서

- ▶ 황선재. 2022. “대전·세종·충청 지역 사회경제적 격차의 추이와 함의.” 『격차와 소외의 관점으로 본 대전·세종 상생 전략』. 대전세종연구원.
- ▶ 최충익. 2008. “패널모형: 시계열 분석과 횡단면 분석을 한번에.” 국토 120-127.
- ▶ 한치록. 2013. “패널자료를 이용한 계량경제분석 방법론.” 『사업체패널조사를 활용한 사업체의 동학 연구』. 한국노동연구원.
- ▶ 전승훈·강성호·임병인. “선형패널자료 분석방법에 관한 비교연구.” 통계연구 9(2): 1-24.
- ▶ Williams, Richard. 2015. “Panel Data: Very Brief Overview.” (<https://www3.nd.edu/~rwilliam/stats2/panel.pdf>)

# # 패널자료 소개

# 자료/데이터 유형

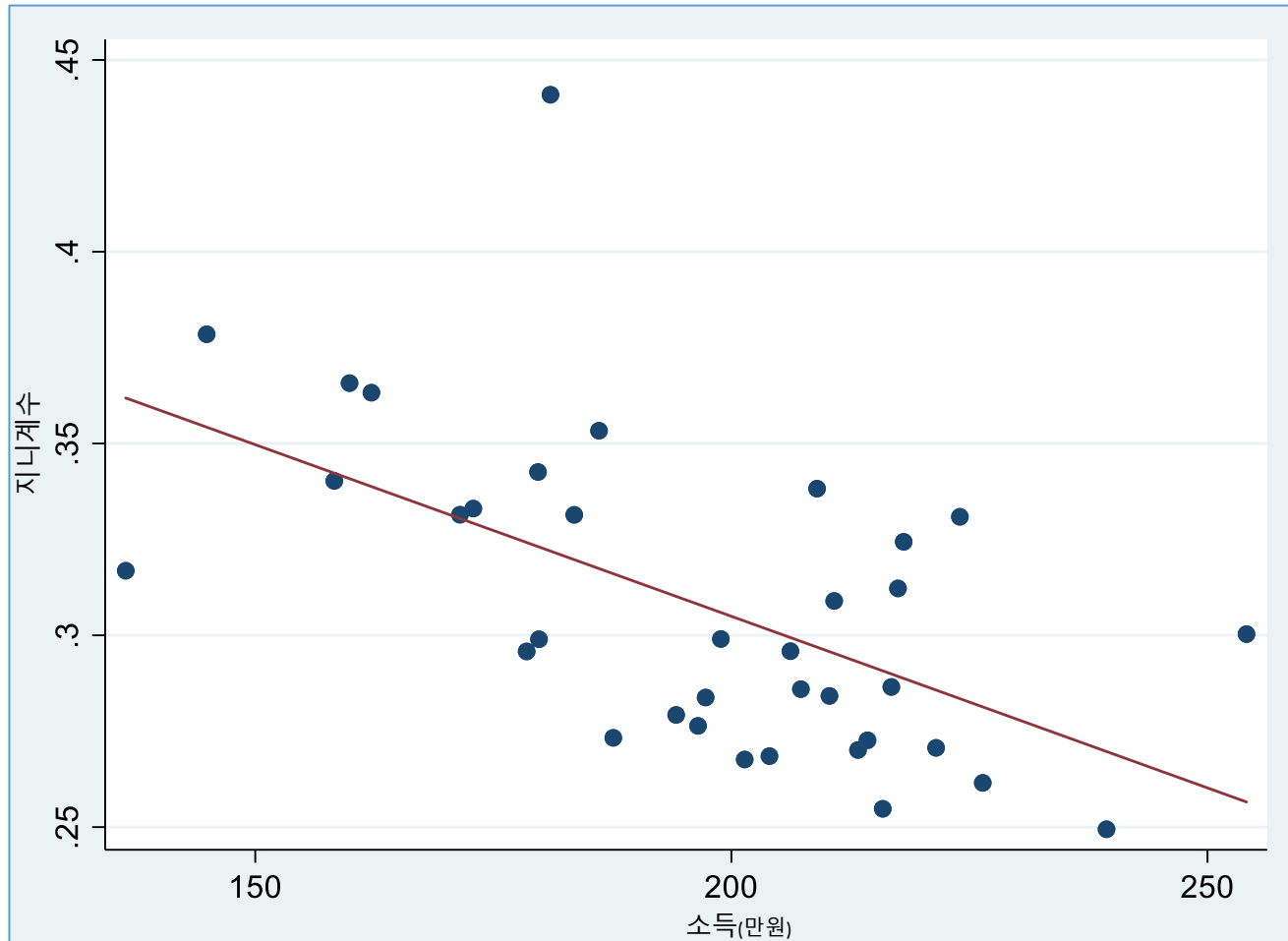
- ▶ **횡단 자료 (cross-sectional data)**
  - ▶ 여러 개체의 특성을 특정 시점에 측정한 자료
    - ▶ 예: 대부분의 일회성 횡단면 서베이 조사
- ▶ **시계열 자료 (time-series data)**
  - ▶ 특정 개체의 특성을 여러 시점에 걸쳐 반복적으로 측정<sup>1</sup>한 자료
    - ▶ 예: 한국의 국내총생산(GDP) 추이
- ▶ **패널/종단 자료 (panel/longitudinal data)**
  - ▶ 여러 개체의 특성을 여러 시점에 걸쳐 반복적으로 측정<sup>2</sup>하되, 매 번 동일한 대상을 측정<sup>3</sup>한 자료. 시계열+횡단자료의 성격
    - ▶ 예: 한국복지패널(KoWePS)
- ▶ **반복/합동 횡단 자료 (repeated/pooled cross-sectional data)**
  - ▶ 여러 개체의 특성을 여러 시점에 걸쳐 반복적으로 측정<sup>4</sup>하되, 매 번 상이한 대상을 측정<sup>5</sup>한 자료
    - ▶ 예: 한국종합사회조사(KGSS)

# 패널자료 예시: long format, unbalanced 자료

	region	year	gini	income_mon	gri_mon
1	SJ	2012	.33144	171.5	.
2	SJ	2013	.31223	217.5	312.9167
3	SJ	2014	.29897	179.8	362.9167
4	SJ	2015	.29586	206.2	326.75
5	SJ	2016	.26152	226.4	317.1667
6	SJ	2017	.24944	239.4	308.9167
7	DJ	2008	.3309	224	142.6333
8	DJ	2009	.29904	198.9	157.7833
9	DJ	2010	.32438	218.1	167.3417
10	DJ	2011	.28653	216.8	179.7833
11	DJ	2012	.28597	207.3	188.5
12	DJ	2013	.2701	213.3	187.925
13	DJ	2014	.2958	178.5	200.1417
14	DJ	2015	.28378	197.3	208.975
15	DJ	2016	.26762	201.4	218.7833
16	DJ	2017	.27264	214.3	231.8833
17	CJ	2008	.33823	209	161.45
18	CJ	2009	.27327	187.6	172.5

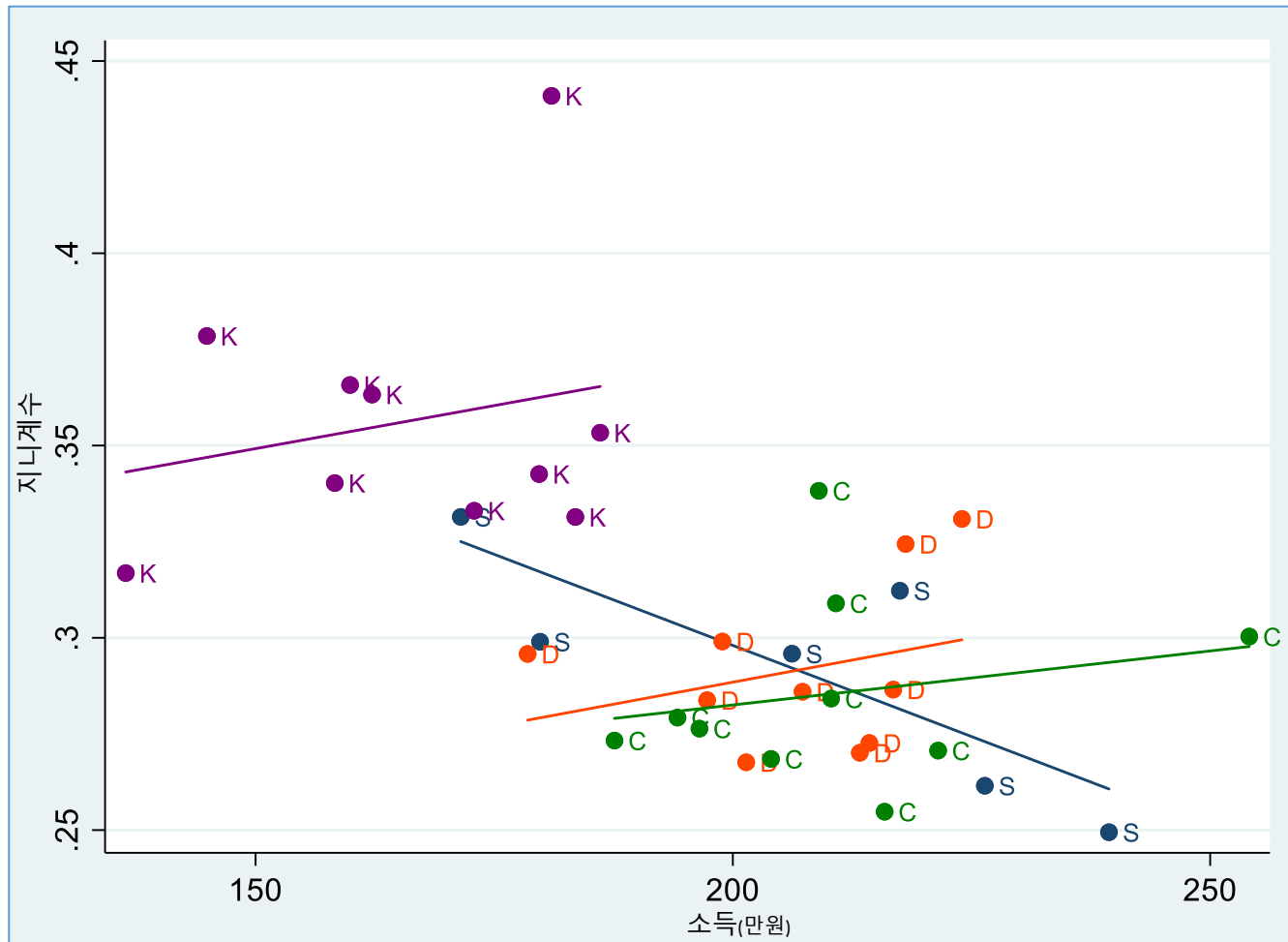
[참조] 황선재(2022)

# 패널자료 분석기법이 필요한 이유: 예시



[참조] 황선재(2022)

# 패널자료 분석기법이 필요한 이유: 예시



[참조] 황선재(2022)



# 패널자료 분석의 장단점

## ▶ 장점

- ▶ 패널자료에는 개체( $i$ ) 정보와 시간( $t$ ) 정보가 동시에 존재함
  - ▶ 예를 들어, 독립변수 값의 변화가 종속변수 값을 변화시키는 정도를 측정하기 위해서 독립변수 값이 상이한 여러 개체들을 비교할 수도 있고(개체간 비교), 동일한 개체 내에서 다양한 시점 간의 차이를 비교할 수도 있음(개체내 비교)
  - ▶ 개체의 정적(static)인 관계만을 추정할 수 있는 횡단자료와는 달리, 개체의 동적(dynamic)인 관계를 추정할 수 있음
- ▶ 개체들의 관찰되지 않는 이질성(unobserved heterogeneity) 요인을 모형에서 고려하여, 누락변수편의(omitted variable bias)를 어느 정도 줄일 수 있음
- ▶ 횡단자료나 시계열자료에 비해 더 많은 정보/표본수와 변수의 변이(variability)를 제공함

## ▶ 단점

- ▶ 수집하기 어려움(특히 표본이탈과 표본선택의 문제)
- ▶ 분석하기 어려움(특히 패널자료분석의 가정(위배)에 대한 이해)

# # 패널자료 분석전략

# 횡단자료 vs. 패널자료 회귀분석

- ▶ 횡단자료 회귀분석 모형

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- ▶ 추정방법: OLS, GLS 등

- ▶ 패널자료 회귀분석 기초 모형 (패널자료 특성/구조 미반영)

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

- ▶ 추정방법: OLS, GLS 등

- ▶ 패널자료 회귀분석 ‘기본 모형’ (총오차=개체특성+고유오차)

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$$

- ▶ 추정방법: OLS, GLS 등

# 회귀분석: 가정

## ▶ OLS 선형 회귀분석의 실질적 가정

### ▶ 1) 모형의 타당성

- ▶ 종속변수와 독립변수의 관계가 제대로 설정되었는가?

### ▶ 2) 독립변수의 외생성

- ▶ 독립변수와 오차항이 서로 상관되지 않는가?
- ▶ OLS 추정량의 **비편향성**(unbiasedness) 및 **일치성**(consistency) 확보

### ▶ 3) 오차항의 비문제성

- ▶ 오차항은 상호 독립추출, 동분산, 정규분포를 만족시키는가?
  - ▶ 독립추출 & 동분산 가정은 OLS 추정량 표집분산의 **효율성**(efficiency)을 위해 필요한 가정 →  $\hat{\beta}_{OLS} = BLUE$
  - ▶ 정규분포 가정은 검정통계량(예. F 검정, t 검정 등) 분포의 정확성을 위해 필요한 가정(but **중심극한정리**에 의해 쉽게 해결됨) →  $\hat{\beta}_{OLS} = BUE$

# 패널자료 회귀분석 기법 종류

- ▶ [기본] 두 가지 ‘기본 모형’
  - ▶ 고정효과 모형 (Fixed-Effects model : **FE**)
    - ▶ FD 추정, WG 추정, LSDV 추정
  - ▶ 임의효과 모형 (Random-Effects model : **RE**)
- ▶ [응용1] 종속변수의 과거값 활용여부에 따라
  - ▶ 정적 패널 모형 & 동적 패널 모형
- ▶ [응용2] 종속변수의 유형에 따라
  - ▶ 선형 회귀 모형(연속형) & 비선형 회귀 모형(범주형)
- ▶ [응용3] 패널자료를 바라보는 관점에 따라
  - ▶ 다층모형(Multi-Level Model: **MLM**)을 적용한 패널자료 분석

# 패널자료 회귀분석 기본전략

- ▶ 기본적으로 패널자료 회귀분석은 횡단 및 시계열 자료 회귀분석과 크게 다르지 않으나, 패널자료의 특징/장점을 활용하는 데이터관리 및 추정방법에 대한 (조금 더 복잡한) 지식이 필요함
- ▶ 예를 들어, 패널자료로 회귀분석을 실시하는 경우, 횡단자료 OLS 추정방법의 가정들(assumptions) 중 어떤 가정이 위배되는지를 잘 생각해보고, 각각의 가정이 위배되었을 때 그 문제를 해결할 수 있는 (조금 더 복잡한) 분석기법을 사용하면 됨
  - ▶ 패널자료라 할지라도 횡단자료 회귀분석의 가정들이 위반되지 않는다면, 그냥 (합동) OLS를 사용하면 됨
- ▶ 위와 같이 ‘조금 더 복잡함’을 감수하면, 회귀분석의 새로운 가능성이 열림을 깨달을 수 있음

# # 패널자료 분석기법: FE 모형

# FE: 정의

## ▶ 선형 고정효과 모형

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$$

## ▶ ‘기본모형’에서와 같이, 총오차를 개체특성과 고유오차의 합으로 두고, 설명변수는 고유오차에 대하여 강외생적이고, 개체특성은 고정효과라 가정하는 모형

### ▶ 고정효과

- ▶ 개체특성이 모형에 포함된 독립변수들과 상관되어 있어서, 추정해야 할 고정(fixed)된 모수로 가정하는 경우
- ▶  $\beta$ 는  $u_i$ 가 고정될 때의 효과를 나타내므로, 고정효과 모형은 기본적으로 개체내 차이로부터 도출되는 함수관계를 정의
- ▶ 고정효과 추정의 핵심은  $u_i$ 를 모형으로부터 통제(제거/추정)하여 설명변수의 내생성 문제를 해결하고  $\beta$ 에 대한 일관된 추정을 하는 것

### ▶ 임의효과

- ▶ 개체특성이 임의(random)로 주어져서, 모형에 포함된 독립변수들과 상관이 되어 있지 않다고 가정하는 경우



# FE: 대표적인 추정방법 세 가지

## ▶ 선형 고정효과 모형 ( $u_i$ 재정렬)

$$y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$$

- ▶ 고정효과 모형은 상수항( $\alpha + u_i$ )이 패널 개체별로 서로 다르면서 고정되어 있다고 가정

## ▶ 대표적인 추정방법

- ▶ 1) First-Difference (FD) 추정
  - ▶  $u_i$  제거하여 개체특성 통제
- ▶ 2) Within-Group (WG) 추정: Stata 기본값
  - ▶  $u_i$  제거하여 개체특성 통제
- ▶ 3) Least Squares Dummy Variable (LSDV) 추정
  - ▶  $u_i$  직접 추정하여 개체특성 통제

# FE: 1) FD 추정방법

## ▶ FE를 FD로 추정하는 법

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it} \quad (1)$$

$$y_{it-1} = \alpha + \beta x_{it-1} + u_i + e_{it-1} \quad (2)$$

$$(y_{it} - y_{it-1}) = \beta(x_{it} - x_{it-1}) + (e_{it} - e_{it-1}) \quad (1)-(2)$$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta e_{it}$$

- ▶  $u_i$ 를 제거하는 방법을 통해 개체특성 효과를 통제
  - ▶  $u_i$ 가 추정과정 상에서 사라지기는 하지만, 여전히 그 효과는 통제되고 있음에 유의!
- ▶  $u_i$ 가 사라지기 때문에  $cov(u_i, x_{it}) \neq 0$  이더라도, POLS 추정을 통해  $\beta$ 에 대한 일치추정량을 구할 수 있음
- ▶ 그러나 그 과정에서 시간불변 변수도 함께 사라지기 때문에 시간불변 변수의 계수를 추정할 수 없다는 단점이 있음

# FE: 1) FD 예시 & 해석

## ▶ FE Model: Regional Gini on equivalized income & GRI per capita

```
. reg D.gini D.income_mon D.gri_mon, nocons // FE model: first-difference estimation
```

Source	SS	df	MS	Number of obs	=	31
Model	.003784408	2	.001892204	F(2, 29)	=	1.70
Residual	.032239924	29	.001111722	Prob > F	=	0.2000
Total	.036024332	31	.001162075	R-squared	=	0.1051
				Adj R-squared	=	0.0433
				Root MSE	=	.03334

D.gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income_mon D1.	.0002965	.0002607	1.14	0.265	-.0002367 .0008297
gri_mon D1.	-.0004027	.0003814	-1.06	0.300	-.0011828 .0003774

## FE: 2) WG 추정방법

### ▶ FE를 WG로 추정하는 법

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it} \quad (1)$$

$$\bar{y}_i = \alpha + \beta \bar{x}_i + u_i + \bar{e}_i \quad (2)$$

$$(y_{it} - \bar{y}_i) = \beta(x_{it} - \bar{x}_i) + (e_{it} - \bar{e}_i) \quad (1)-(2)$$

- ▶  $u_i$ 를 제거하는 방법을 통해 개체특성 효과를 통제
  - ▶  $u_i$ 가 추정과정 상에서 사라지기는 하지만, 여전히 그 효과는 통제되고 있음에 유의!
- ▶  $u_i$ 가 사라지기 때문에  $cov(u_i, x_{it}) \neq 0$  이더라도, POLS 추정을 통해  $\beta$ 에 대한 일치추정량을 구할 수 있음
- ▶ 그러나 그 과정에서 시간불변 변수도 함께 사라지기 때문에 시간불변 변수의 계수를 추정할 수 없다는 단점이 있음

# FE: 2) WG 예시 & 해석

## ▶ FE Model: Regional Gini on equivalized income & GRI per capita

```
. xtreg gini income_mon gri_mon, fe // FE model: within-group estimation
```

Fixed-effects (within) regression  
Group variable: region

Number of obs = 35  
Number of groups = 4

R-sq:

within = 0.2109  
between = 0.0396  
overall = 0.0130

Obs per group:

min = 5  
avg = 8.8  
max = 10

corr(u\_i, Xb) = -0.6322

F(2,29) = 3.88  
Prob > F = 0.0322

gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income_mon	1.68e-06	.0002642	0.01	0.995	-.0005387	.000542
gri_mon	-.0003851	.0001402	-2.75	0.010	-.0006719	-.0000984
_cons	.3988219	.0664566	6.00	0.000	.2629028	.534741
sigma_u	.04637788					
sigma_e	.02518505					
rho	.7722648	(fraction of variance due to u_i)				

F test that all u\_i=0: F(3, 29) = 10.92

Prob > F = 0.0001

## FE: 3) LSDV 추정방법

### ▶ FE를 LSDV로 추정하는 법

$$y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$$

$$y_{it} = \sum_{i=1}^n \alpha_i + \beta x_{it} + e_{it} \quad (\text{where } \alpha_i = \alpha + u_i)$$

- ▶  $u_i$ 를 직접추정하는 방법을 통해 개체특성 효과를 통제
  - ▶ 실제로 추정되는 값은  $u_i$ 가 아니라  $\alpha_i$
  - ▶ 시간에 따라 변하지 않은 개체 특성은 모두  $\alpha_i$ 에 포함되어 추정됨(따라서 시간불변 변수 효과는 별도로 추정 안됨)
- ▶ LSDV 추정 결과는 WG 추정 결과와 완전히 일치

# FE: 3) LSDV 예시 & 해석

## ▶ FE Model: Regional Gini on equivalized income & GRI per capita

```
. xi: reg gini income_mon gri_mon i.region // fixed-effect model: LSDV estimat
> ion - ui
i.region          _Iregion_0-3          (naturally coded; _Iregion_0 omitted)
```

Source	SS	df	MS	Number of obs	=	35
Model	.039226508	5	.007845302	F(5, 29)	=	12.37
Residual	.018394321	29	.000634287	Prob > F	=	0.0000
Total	.057620829	34	.00169473	R-squared	=	0.6808
				Adj R-squared	=	0.6257
				Root MSE	=	.02519

gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income_mon	1.68e-06	.0002642	0.01	0.995	-.0005387	.000542
gri_mon	-.0003851	.0001402	-2.75	0.010	-.0006719	-.0000984
_Iregion_1	-.0448152	.0239944	-1.87	0.072	-.0938892	.0042589
_Iregion_2	-.0432832	.0215878	-2.00	0.054	-.0874352	.0008689
_Iregion_3	.0535636	.0206362	2.60	0.015	.0113578	.0957693
_cons	.408689	.0789602	5.18	0.000	.2471973	.5701807

# FE: FD, WG, LSDV 추정값 비교

```
. quietly reg D.gini D.income_mon D.gri_mon, nocons
.           estimates store FD
. quietly xtreg gini income_mon gri_mon, fe
.           estimates store WG
. quietly xi: reg gini income_mon gri_mon i.region
.           estimates store LSDV
. estimates table FD WG LSDV, b(%9.6f) star(0.1 0.05 0.001)
```

Variable	FD	WG	LSDV
income_mon			
D1.	0.000296	0.000002	0.000002
--.			
gri_mon			
D1.	-0.000403	-0.000385**	-0.000385**
--.			
_Iregion_1			-0.044815*
_Iregion_2			-0.043283*
_Iregion_3			0.053564**
_cons		0.398822***	0.408689***

legend: 2\* p<.1; \*\* p<.05; \*\*\* p<.001



# # 패널자료 분석기법: RE 모형

# RE: 정의

## ▶ 선형 임의효과 모형

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$$

- ▶ ‘기본모형’에서와 같이, 총오차를 개체특성과 고유오차의 합으로 두고, 설명변수는 고유오차에 대하여 강외생적이고, 개체특성은 임의효과라 가정하는 모형
  - ▶ 임의효과
    - ▶ 개체특성이 **임의(random)**로 주어져서, 모형에 포함된 독립변수들과 상관이 되어 있지 않다고 가정하는 경우
  - ▶ 고정효과
    - ▶ 개체특성이 모형에 포함된 독립변수들과 상관되어 있어서, 추정해야 할 고정(fixed)된 모수로 가정하는 경우
- ▶ 상기의 식을 그냥 POLS로 추정하면 총오차의 자기상관 때문에 효율적이지 않으므로, ‘특별한 변환’을 하여 추정해야함(GLS)

# RE: 추정방법

## ▶ RE 추정방법 & 'theta( $\theta$ )'

$$(y_{it} - \theta_i \bar{y}_i) = (\alpha - \theta_i \alpha) + (\beta x_{it} - \beta \theta_i \bar{x}_i) + [(u_i - \theta_i u_i) + (e_{it} - \theta_i \bar{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_u^2 + \sigma_e^2}}$$

- ▶ 여기서  $\theta$  는 변환된 오차항  $((u_i - \theta_i u_i) + (e_{it} - \theta_i \bar{e}_i))$  이 '기본가정' 하에서 등분산적이고 자기상관이 없도록 선택됨
- ▶  $\theta$  는 그 정의상 0에서 1사이의 값을 갖게 되는데, 0에 가까울 수록 RE 추정값은 POLS 추정값에 가깝게 되고, 1에 가까울 수록 FE 추정값에 가깝게 됨

# RE: 예시 & 해석

## ▶ RE Model: Regional Gini on equivalized income & GRI per capita

```
. xtreg gini income_mon gri_mon, re theta // RE model
```

```
Random-effects GLS regression
Group variable: region
```

```
Number of obs   =      35
Number of groups =       4
```

```
R-sq:
  within = 0.0012
  between = 0.9694
  overall = 0.3202
```

```
Obs per group:
  min =      5
  avg =     8.8
  max =     10
```

```
corr(u_i, X) = 0 (assumed)
```

```
Wald chi2(2) =     15.07
Prob > chi2   =     0.0005
```

theta				
min	5%	median	95%	max
0.0000	0.0000	0.0000	0.0000	0.0000

gini	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
income_mon	-.0009234	.0002427	-3.80	0.000	-.0013991	-.0004476
gri_mon	-.0000433	.0001065	-0.41	0.684	-.000252	.0001654
_cons	.499861	.0608897	8.21	0.000	.3805193	.6192027
sigma_u	0					
sigma_e	.02518505					
rho	0	(fraction of variance due to u_i)				

# FE or RE?

## ▶ 하우스만 검정(The Hausman specification test)

$$H_0: cov(x_{it}, u_i) = 0$$

$$H_a: cov(x_{it}, u_i) \neq 0$$

### ▶ 귀무가설을 기각하지 못할 경우

- ▶ RE 추정량과 FE 추정량 모두 일치추정량이나 RE 추정량이 더 효율적이므로 RE 모형 선택

### ▶ 귀무가설을 기각할 경우

- ▶ RE 추정량은 일치추정량이 아니므로, 일치추정량인 FE 모형을 선택

# FE or RE?

```
. quietly xtreg gini income_mon gri_mon, fe
.           estimates store FE
. quietly xtreg gini income_mon gri_mon, re
.           estimates store RE
. hausman FE RE
```

	—— Coefficients ——			
	(b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
income_mon	1.68e-06	-.0009234	.0009251	.0001043
gri_mon	-.0003851	-.0000433	-.0003418	.0000912

b = consistent under Ho and Ha; obtained from xtreg  
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

```
chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
        =      83.01
Prob>chi2 =      0.0000
```

감사합니다 :)