패널데이터 분석 방법론

2021 KHP 학술대회 특강

Dec 10, 2021

Table of Contents

smart_khp 패키지

- 1기 의료패널 데이터: 2008년 ~ 2018년(11개년) 패널데이터

  2기 의료패널 데이터: 2019년(1개년) 횡단면 데이터

- 1기 KHP: 12개 부문 ($12 \times 11 = 132$개 data files)

  2기 KHP: 4개 부문(hh, ind, ms, phi)

- KHP 1기 데이터를 손쉽게 패널데이터로 만들 수 있는 Stata 명령어 패키지: smart_khp_v4

smart_khp 패키지 (Cont'd)

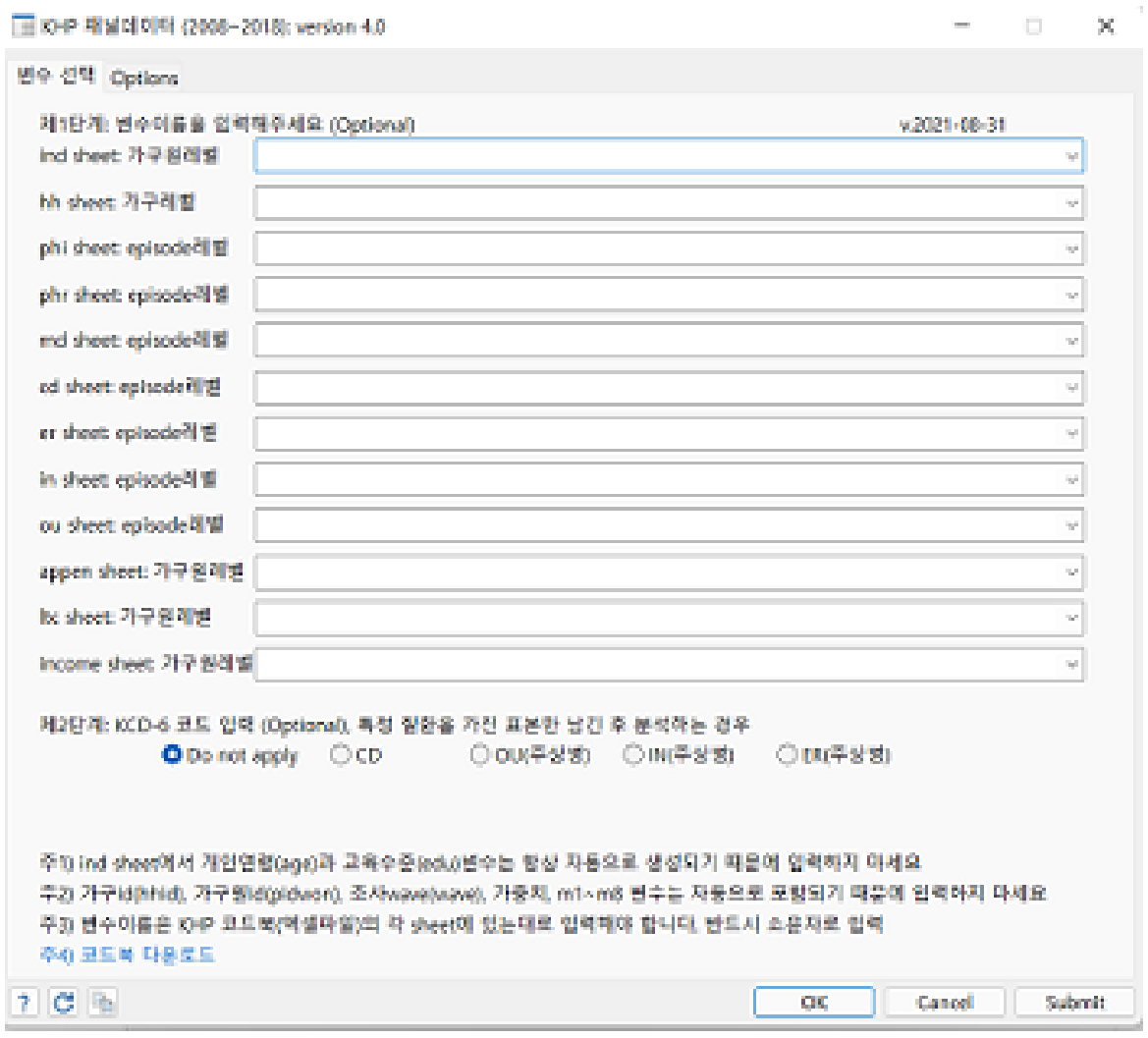
- 코드북에 있는 변수 이름을 그대로 입력 또는 선택
- 특정 질환(KCD)을 가진 표본만 선택 가능(OU, IN, ER, CD)

- 패널데이터로 만들고자 하는 year를 선택
- Stata 명령문으로 실행

```
smart_khp_v4 , ind(c3) appen(s2 s17) ///
ou_kcd kcd(K2 K30 K31) wave(2012-2018)
```

note) `fid( )` 옵션 required

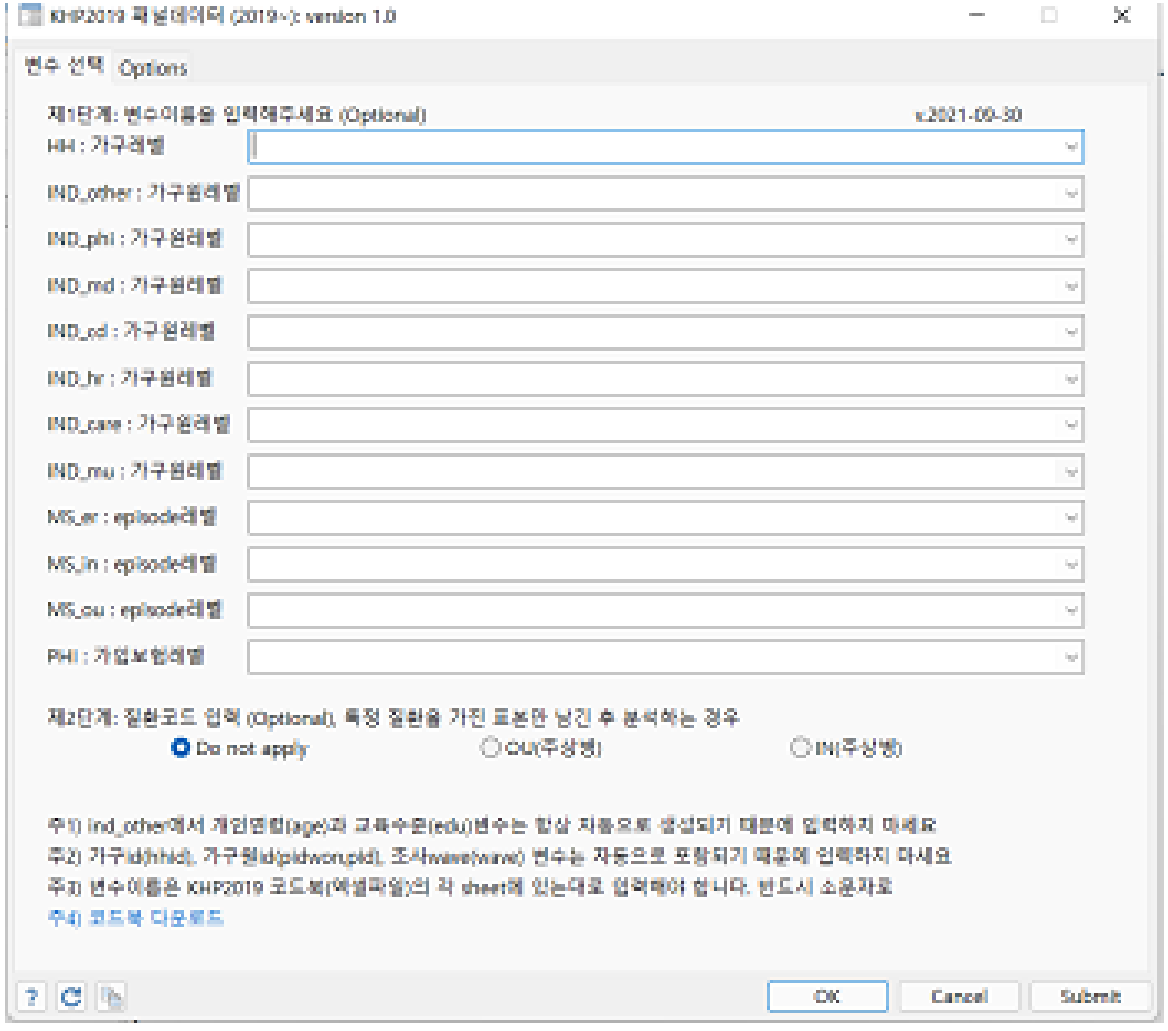
smart_khp 패키지 (Cont'd)



smart_khp 패키지 (Cont'd)

- KHP 2기 데이터를 유사한 방식으로 활용할 수 있는 smart_khp2019 명령어 개발
- 코드북의 HH, IND, MS, PHI sheet에 있는 변수이름을 그대로 입력
- OU와 IN에서 질환명을 선택하면 해당 질환으로 치료받은 표본만 선택함

smart_khp 패키지 (Cont'd)



smart_khp 패키지 (Cont'd)

- 참고문헌
  1. 민인식(2021), 한국의료패널(KHP) 활용을 위한 Stata 패키지 개발, *의료경영학 연구*, 15(1), 25-35
  2. 민인식 교수 블로그: blog.naver.com/housingdata

DID 추정: two-period 패널

- Impact Evaluation: 정책 수혜자의 성과(outcome)가 해당 정책프로그램에 기인한 것인지 다른 요인때문인 분석하는 것이 목적

  - 단순히 정책 프로그램 전후(before-after) 차이를 비교하는 것이 아니라고 causality를 추정

  - DID는 특정 프로그램(정책)에 참여하였을 성과에 미치는 인과관계를 식별하는데 가장 자주 쓰이는 방법론

DID 추정: two-period 패널 (Cont'd)

- 평균 처리효과 정의
  Average Treatment Effect(ATE) for all i

$$= E(Y_{1i} - Y_{0i}) = E(Y_{1i}) - E(Y_{0i})$$

- 특정 표본 i에게 동시에 treated outcome과 non-treated outcome이 발생할 수 없다. $\Longrightarrow$ missing data problem

- Y_{0i}를 대신하는 counterfactual outcome을 찾아야 한다.

DID 추정: two-period 패널 (Cont'd)

- two-period 패널데이터에서 pre-treatment(0 시점)과 post-treatment(1 시점) 그리고 treatment와 control 그룹으로 구분 가능해야 함

	Pre	Post
Treatment	Y_0^T	Y_1^T
Control	Y_0^C	Y_1^C

- pre-treatment: 모든 표본이 treatment 적용을 받지 않음.
  post-treatment: control 그룹은 여전히 treatment를 받지 않고 treatment 그룹만 정책 수혜의 대상

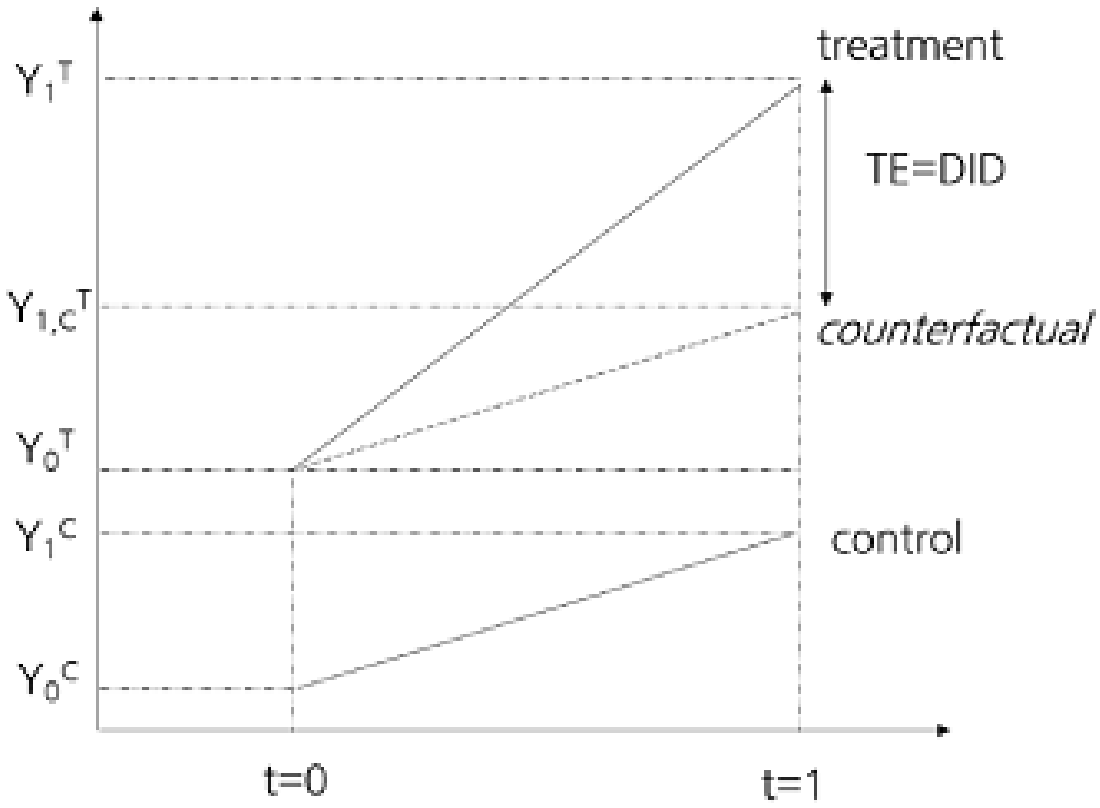
DID 추정: two-period 패널 (Cont'd)

- Q) reflective comparison: Y_0^T와 Y_1^T의 차이를 비교

- Q) cross-sectional comparison: Y_1^T과 Y_1^C의 차이를 비교

  Selection bias: $E(Y_{0i}|T_i = 1) - E(Y_{0i}|T_i = 0)$

DID 추정: two-period 패널 (Cont'd)



- t=0 시점에서 treatment와 control 그룹의 차이는 (정책이 시행되지 않았다면) t=1 시점에서도 그대로 유지
  $\Longrightarrow$ Parallel-trend 가정

DID 추정: two-period 패널 (Cont'd)

- DID 추정량

$$ATE_{DID} = E(Y_1^T - Y_0^T) - E(Y_1^C - Y_0^C)$$

- DID 추정량: selection bias(선택편의) 원인을 두 그룹의 시간불변 이질성(time-invariant heterogeneity)이라고 가정
- 패널데이터에서 시간불변 이질성(group fixed-effects)을 통제하여 ATE를 얻는다.
- PSM과 비교하여 장점: selection only on observable X 가정을 완화

DID 추정: two-period 패널 (Cont'd)

- DID 추정을 위한 Pooling Regression

$$y_i = \alpha + \beta t + \gamma T_i + \delta(t \times T_i) + e_i \quad (1)$$

where t time dummy, T_i treatment group dummy

$$\delta = [E(Y_1^T) - E(Y_0^T)] - [E(Y_1^C) - E(Y_0^C)]$$

- Panel Regression : Fixed-effects 추정

$$y_{it} = \alpha + \beta T_{it} + \theta t + u_i + \epsilon_{it} \quad (2)$$

$$\widehat{\beta}_{FE} = \widehat{\delta}_{(1)}$$

DID 추정: two-period 패널 (Cont'd)

- Stata 실습 : Pooling Regression

time: 2016=0, 2018=1

	2016	2018
Treatment	No USC	USC
Control	No USC	No USC

$$m\_exp = \alpha + \beta D_{2018} + \gamma D_{Treat} + \delta(D_{2018} \times Treat) + e$$

DID 추정: two-period 패널 (Cont'd)

```
. reg ly i.time i.treat i.time#i.treat
```

Source	SS	df	MS	Number of obs	=	9,650
				F(3, 9646)	=	120.06
Model	884.178476	3	294.726159	Prob > F	=	0.0000
Residual	23680.1392	9,646	2.45491802	R-squared	=	0.0360
				Adj R-squared	=	0.0357
Total	24564.3177	9,649	2.54578896	Root MSE	=	1.5668

ly	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.time	.2771028	.0405634	6.83	0.000	.1975901	.3566155
1.treat	.601134	.0464345	12.95	0.000	.5101126	.6921554
time#treat						
1 1	-.0698212	.0656683	-1.06	0.288	-.1985449	.0589024
_cons	12.17074	.0286826	424.32	0.000	12.11452	12.22697

DID 추정: two-period 패널 (Cont'd)

- Stata 실습: Panel Regression

$$m\_exp = \alpha + \beta D_{2018} + \gamma D_{Treat} + \delta(D_{2018} \times Treat) + e$$

- Note) Panel Regression에서 $Treat_{it}$와 Pooling Regression에서 $Treat_i$는 서로 차이가 있다.

DID 추정: two-period 패널 (Cont'd)

```
. tsset pidwon time
Panel variable: pidwon (strongly balanced)
 Time variable: time, 0 to 1
         Delta: 1 unit
. xtreg ly i.time i.usc , fe
Fixed-effects (within) regression               Number of obs      =      9,650
Group variable: pidwon                          Number of groups   =      4,825
R-squared:                                      Obs per group:
     Within  = 0.0202                                         min =          2
     Between = 0.0431                                         avg =        2.0
     Overall = 0.0031                                         max =          2
                                                F(2,4823)          =      49.68
corr(u_i, Xb) = -0.0292                         Prob > F           =     0.0000
```

ly	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.time	.2771028	.0322418	8.59	0.000	.2138941	.3403115
1.usc	-.0698212	.0521965	-1.34	0.181	-.1721502	.0325077
_cons	12.40011	.017929	691.62	0.000	12.36496	12.43526
sigma_u	1.328294					
sigma_e	1.2453862					
rho	.53218034	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(4824, 4823) = 2.22                     Prob > F = 0.0000
```

DID 추정: two-period 패널 (Cont'd)

- Stata 실습: 17버전 명령어

```
. didregress (ly ) (usc), time(time) group(pidwon)
Number of groups and treatment time
Time variable: time
Control:       usc = 0
Treatment:     usc = 1
```

	Control	Treatment
Group		
pidwon	2984	1841
Time		
Minimum	0	1
Maximum	0	1

```
Difference-in-differences regression                 Number of obs = 9,650
Data type: Repeated cross-sectional
                              (Std. err. adjusted for 4,825 clusters in pidwon)
```

ly	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
usc						
(1 vs 0)	-.0698212	.0703056	-0.99	0.321	-.2076523	.0680098

```
Note: ATET estimate adjusted for group effects and time effects.
```

DID 추정: two-period 패널 (Cont'd)

Note) `didregress` 명령문에서 **usc** 변수는 $Treat_{it}$ 변수와 같다.

Q) 추가적인 통제변수의 포함: 시간가변 변수(time-varying) vs. 시간불변 변수(time-invariant)

DID 추정: Multi-period 패널

- Two-period 패널 대신 Multi-period(가령 10개 시점) 패널데이터가 주어졌다고 가정

	Pre (1 ∼ 5)	Post (6 ∼ 10)
Treatment	Y_0^T	Y_1^T
Control	Y_0^C	Y_1^C

- Pre와 Post treatment 기간의 성과(outcome)를 Pooling 하여 $\widehat{\delta}_{DID}$를 얻는 아이디어 : Homogeneous ATE for each time period

$$Y_{it} = \gamma_t + \gamma_i + \beta X_{it} + \delta T_{it} + \epsilon_{it}$$

Note) Post-treatment의 각 시점(each time period)에서 DID 추정량을 얻는 것은 아니다.

DID 추정: Multi-period 패널 (Cont'd)

- Panel Regression with multiple time dummies : Two-way FE

$$y_{it} = \alpha + \sum_{t=2}^{10} \beta_t time + \delta T_{it} + u_i + \epsilon_{it}$$

where T_{it}는 각 시점(at time t)에서 treated 여부를 나타내는 더미변수

DID 추정: Multi-period 패널 (Cont'd)

- Parallel Trend 가설검정: Pre-treatment 기간 동안 treated 그룹과 control 그룹 간 시간에 따른 성장곡선에서 기울기 차이가 있는지 검정

H_0 : **Linear trends are parallel for pre-treatment period**

- Stata에서 가설검정 결과와 Parallel linear trend를 확인할 수 있는 그래프를 제시

Note) two-period panel 데이터에서는 parallel trend assumption 검정이 불가

DID 추정: Multi-period 패널 (Cont'd)

```
. xtdidregress (y1 ) (treated1), group(id1) time(t1)
Number of groups and treatment time
Time variable: t1
Control:       treated1 = 0
Treatment:     treated1 = 1
```

	Control	Treatment
Group		
id1	102	98
Time		
Minimum	1	6
Maximum	1	6

```
Difference-in-differences regression                 Number of obs = 2,000
Data type: Longitudinal
                                        (Std. err. adjusted for 200 clusters in id1)
```

y1	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
treated1						
(Treated vs Untreated)	.4825449	.0275446	17.52	0.000	.4282281	.5368616

```
Note: ATET estimate adjusted for panel effects and time effects.
```

DID 추정: Multi-period 패널 (Cont'd)

```
. xtreg y1 i.t1 i.treated1 , fe
```

```
Fixed-effects (within) regression               Number of obs      =       2,000
Group variable: id1                             Number of groups   =         200

R-squared:                                      Obs per group:
     Within  = 0.3909                                         min =          10
     Between = 0.5151                                         avg =        10.0
     Overall = 0.4038                                         max =          10

                                                F(10,1790)         =      114.89
corr(u_i, Xb) = 0.1822                          Prob > F           =      0.0000
```

y1	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
t1						
2	-.1488766	.0337923	-4.41	0.000	-.215153	-.0826001
3	-.1826114	.0337923	-5.40	0.000	-.2488878	-.116335
4	-.3780333	.0337923	-11.19	0.000	-.4443097	-.3117568
5	-.4852541	.0337923	-14.36	0.000	-.5515306	-.4189777
6	-.8066861	.0368964	-21.86	0.000	-.8790506	-.7343215
7	-.7912178	.0368964	-21.44	0.000	-.8635823	-.7188532
8	-.8238595	.0368964	-22.33	0.000	-.8962241	-.751495
9	-.9005149	.0368964	-24.41	0.000	-.9728795	-.8281504
10	-.9213645	.0368964	-24.97	0.000	-.9937291	-.849
treated1						
Treated	.4825449	.0302308	15.96	0.000	.4232536	.5418362
_cons	7.829716	.0238947	327.68	0.000	7.782851	7.87658
sigma_u	.23988594					
sigma_e	.33792264					
rho	.3350779	(fraction of variance due to u_i)				

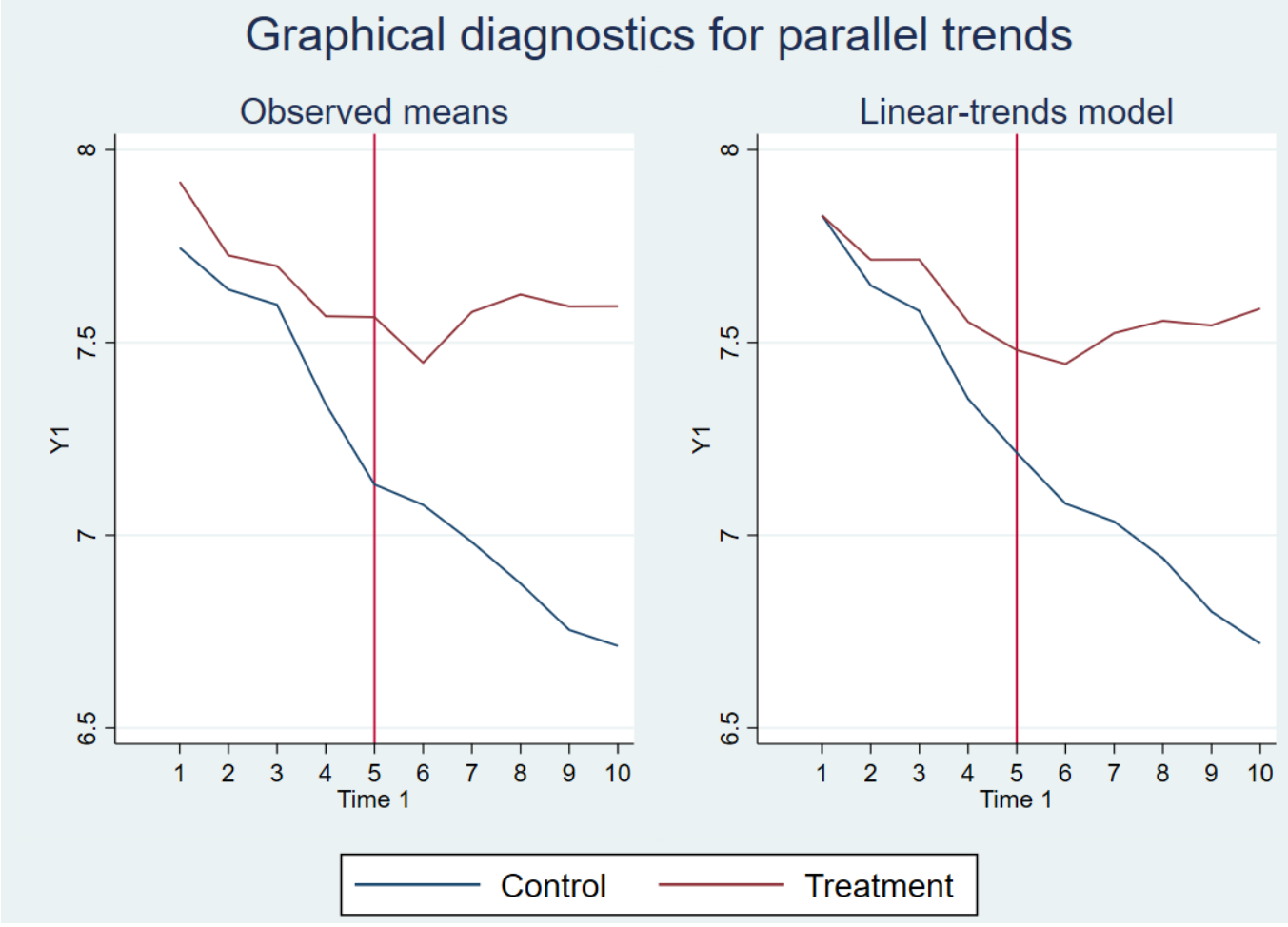
```
F test that all u_i=0: F(199, 1790) = 4.58                     Prob > F = 0.0000
```

DID 추정: Multi-period 패널 (Cont'd)

```
. estat ptrends
Parallel-trends test (pretreatment time period)
H0: Linear trends are parallel
F(1, 199) =  19.75
 Prob > F = 0.0000
```

Q) 가설검정 결과 해석:

DID 추정: Multi-period 패널 (Cont'd)



DID 추정: Multi-period 패널 (Cont'd)

- each time period에서 heterogeneous *ATE* 를 추정하고자 하는 경우, 아래 책의 1장 참고

  민인식.최필선(2021), "STATA 고급통계분석(16-17버전)", 2판, 지필출판사

Latent Growth Model

- 패널데이터에서 Within subject(individual)의 시간에 따른 변화 패턴을 추정할 때 Longitudinal LGM을 설정

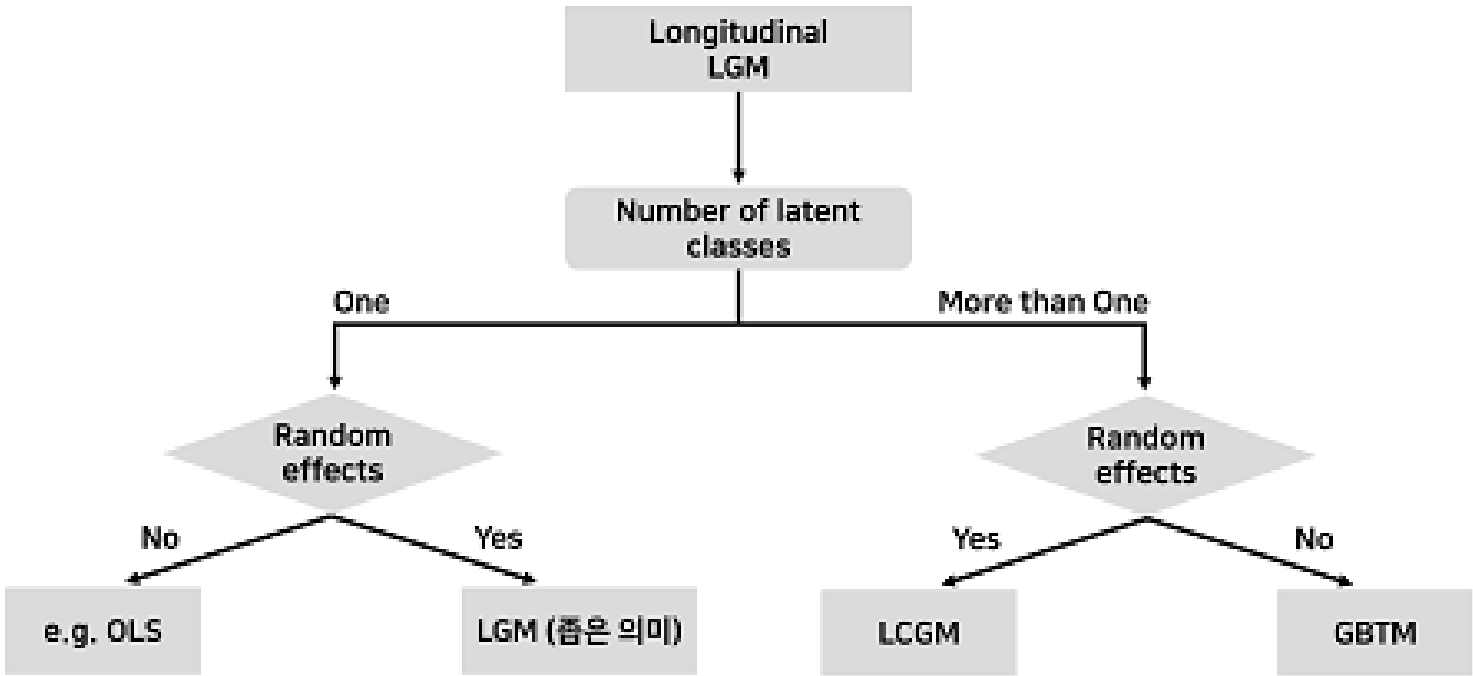
  - 물론 inter-subject의 시간에 따른 변화 패턴을 추정하는 것도 포함

- 시간에 따른 변화: time trend, time path, growth curve, latent trajectory

- 결혼 이후 혼인지속기간에 따른 부부만족도의 변화
  중학교 ~ 고등학교 시점까지 청소년 비행 변화
  고령자의 은퇴이후 시간에 따른 건강상태 변화 등

Latent Growth Model (Cont'd)

- Family of Longitudinal LGM



- LCGM: Latent Class Growth Model
  GBTM: Group-Based Trajectory Model

Latent Growth Model (Cont'd)

- (좁은 의미) LGM: single class about the growth pattern & individual growth heterogeneity (individual trajectories)
- data requirement: panel data (repeated subjects)

  - at least three waves of data : time period가 많을수록 다양한 성장패턴을 가정

  - linear growth, curvilinear growth, cubic growth 등

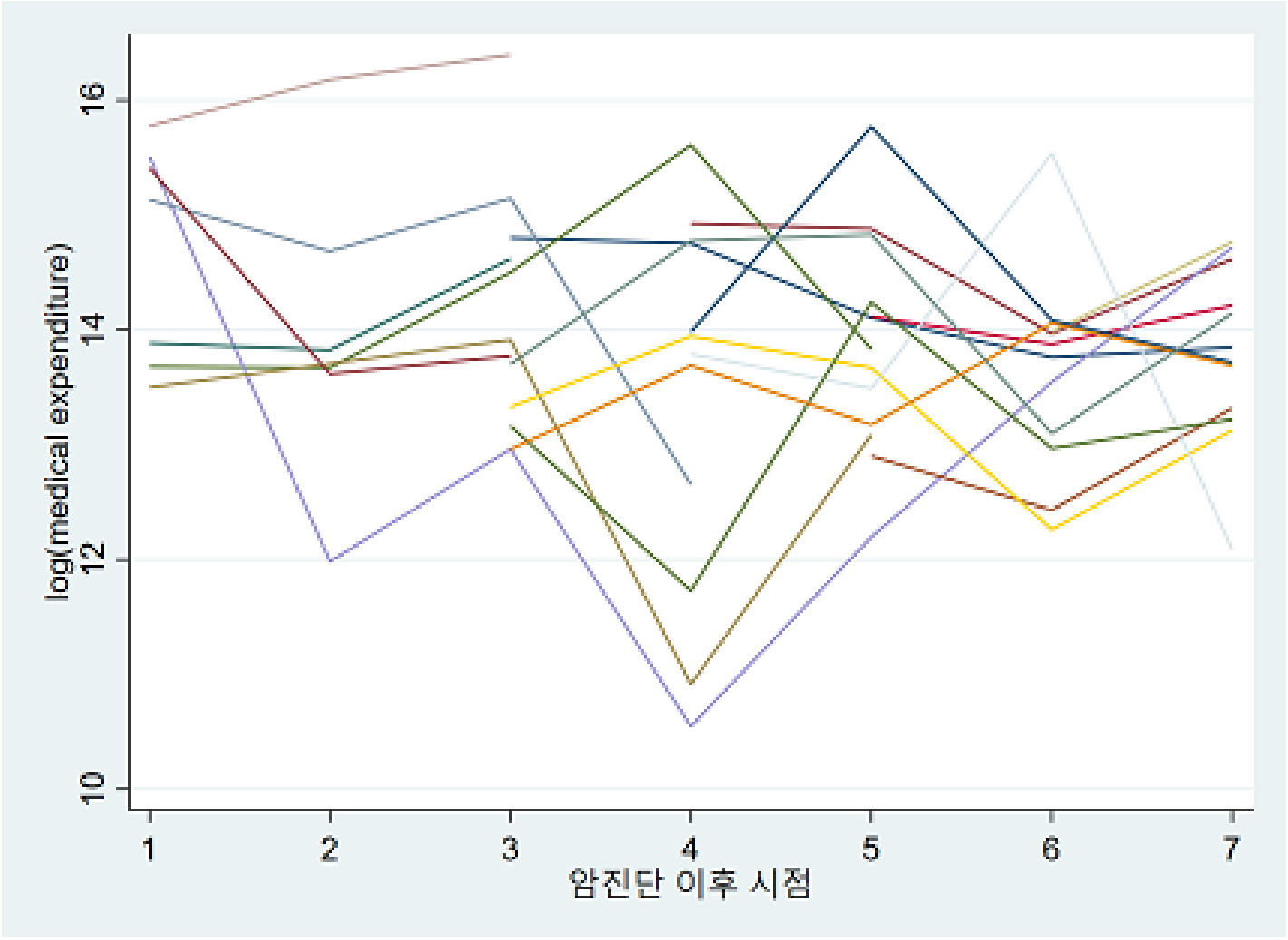
- 단점: time 변수와 종속변수(outcome) 간 growth pattern에 대한 theoretical background가 없음. empirically functional form을 찾아야 한다.

Latent Growth Model (Cont'd)

- 만성질환 중 암진단 이후 시점(time)과 종속변수(log(medical expenditure))의 관계
- 암 진단 시점에서 40세 이상인 환자만 선택
- 암 진단 이후 의료비 성장패턴이 개인별로 heterogeneous
  $\Longrightarrow$ individual trajectory 가정

Latent Growth Model (Cont'd)

- 무작위로 선택된 20명의 의료비 성장패턴



Latent Growth Model (Cont'd)

LGM: Model Specifications

- Traditional Regression : No random effects

$$y_{ij} = \beta_0 + \beta_1 Time_{ij} + e_{ij} \quad \textbf{Model 1}$$

where j individual id and i time period id

가정: $e_{ij} \sim N(0, \sigma_e^2)$

Note) 패널 그룹 j가 서로 같은 time period를 가질 필요는 없음

Q1) β_1 해석:

LGM: Model Specifications (Cont'd)

- Linear Growth with random intercept

$$y_{ij} = \beta_{0j} + \beta_1 Time_{ij} + e_{ij} \quad \textbf{Model 2}$$

$$\text{상위 레벨 : } \beta_{0j} = \gamma_0 + u_j$$

Full model:

$$y_{ij} = \gamma_0 + \beta_1 Time_{ij} + u_j + e_{ij}$$

가정: $u_j \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2), cov(u_j, e_{ij}) = 0$

LGM: Model Specifications (Cont'd)

Q2) Model 2에서 growth pattern 가정은 무엇인가?

Q3) γ_0와 β_1에 대한 해석은?

LGM: Model Specifications (Cont'd)

- Linear Growth with random intercept and slope

$$y_{ij} = \beta_{0j} + \beta_{1j} Time_{ij} + e_{ij} \quad \textbf{Model 3}$$

$$\text{상위레벨} : \beta_{0j} = \gamma_0 + u_{0j}$$

$$\text{상위레벨} : \beta_{1j} = \gamma_1 + u_{1j}$$

Full model:

$$y_{ij} = \gamma_0 + \gamma_1 Time_{ij} + u_{0j} + u_{1j} Time_{ij} + e_{ij}$$

가정: $u_{0j} \sim N(0, \sigma^2_{u0}), u_{1j} \sim N(0, \sigma^2_{u1}), e_{ij} \sim N(0, \sigma^2_e)$
$cov(u_{0j}, e_{ij}) = cov(u_{1j}, e_{ij}) = 0$: **cross-level covariances=0**
$cov(u_{0j}, u_{1j})$ may not be zero: unstructured covariance

LGM: Model Specifications (Cont'd)

Q4) Model 3에서 growth pattern 가정은 무엇인가?

Q5) Model 3에서 individual trajectory는 어떻게 표현할 수 있는가?

Q6) Model 3에서 $u_{1j} > 0$ 의미는?

LGM: Model Specifications (Cont'd)

- curvilinear growth with random effects

$$y_{ij} = \beta_{0j} + \beta_{1j} Time_{ij} + \beta_{2j} Time_{ij}^2 + e_{ij} \quad \textbf{Model 4}$$

$$\text{상위레벨} : \beta_{0j} = \gamma_0 + u_{0j}$$

$$\text{상위레벨} : \beta_{1j} = \gamma_1 + u_{1j}$$

$$\text{상위레벨} : \beta_{2j} = \gamma_2 + u_{2j}$$

Full model:

$$y_{ij} = \gamma_0 + \gamma_1 Time_{ij} + \gamma_2 Time_{ij}^2$$
$$+ u_{0j} + u_{1j} Time_{ij} + u_{2j} Time_{ij}^2 + e_{ij}$$

가정: cross-level covariances are all zeroes
same-level covariances may not be zero (unstructured)

LGM: Model Specifications (Cont'd)

Q7) Model 4에서 growth pattern 가정은 무엇인가?

Q8) Model 4에서 $\gamma_2 < 0$이고 $\gamma_1 < 0$ 의미는 무엇인가?

LGM: Model Specifications (Cont'd)

- Same time effects for known sub-groups

$$y_{ij} = \beta_{0j} + \beta_{1j} Time_{ij} + e_{ij} \qquad \textbf{Model 5}$$

$$\text{상위레벨} : \beta_{0j} = \gamma_0 + \delta_0 Male_j + u_{0j}$$

$$\text{상위레벨} : \beta_{1j} = \gamma_1 + u_{1j}$$

Full model:

$$y_{ij} = \gamma_0 + \delta_0 Male_j + \gamma_1 Time_{ij} + u_{0j} + u_{1j} Time_{ij} + e_{ij}$$

- 개인간 성장 기울기(growth slope)는 서브그룹(남녀) 간 차이가 없다.
- 오차항에 대한 가정은 Model 3과 같다.

LGM: Model Specifications (Cont'd)

Q9) Model 5에서 $\delta_0 > 0$ 의미는 무엇인가?

LGM: Model Specifications (Cont'd)

- Different time effects for known sub-groups

$$y_{ij} = \beta_{0j} + \beta_{1j} Time_{ij} + e_{ij} \qquad \textbf{Model 6}$$

$$\text{상위레벨} : \beta_{0j} = \gamma_0 + \delta_0 Male_j + u_{0j}$$

$$\text{상위레벨} : \beta_{1j} = \gamma_1 + \delta_1 Male_j + u_{1j}$$

Full model:

$$y_{ij} = \gamma_0 + \delta_0 Male_j + \gamma_1 Time_{ij}$$
$$+\delta_1 (Male_j \times Time_{ij}) + u_{0j} + u_{1j} Time_{ij} + e_{ij}$$

- 개인간 성장 기울기(growth slope)는 그룹(남녀) 간 차이가 발생한다.
- 오차항에 대한 가정은 Model 4과 같다.

LGM: Model Specifications (Cont'd)

Q10) Model 6에서 $\delta_1 > 0$ 의미는 무엇인가?

Q11) Model 6에서 남자에 해당하는 개인 j의 평균적 기울기는 어떻게 표현할 수 있나요?

Q12) Model 6에서 남자에 해당하는 개인 j의 이질적 기울기는 어떻게 표현할 수 있나요?

LGM: More Advanced Topics

- 개인 j가 성별에 따른 평균적 성장에서 차이가 있는지
  $\Longrightarrow$ Model 6을 통해 가설검정
- 개인 j가 성별에 따른 잠재성장(latent growth)에서 차이가 있는지
  $\Longrightarrow$ random effects에 대해 성별에 따른 이질성 허용
- random slope에 해당하는 u_{1j}의 분산동일성 가설검정

$$H_0 : var(u_{1j})_{male} = var(u_{1j})_{female}$$

- 민인식.최필선(2021), "STATA 고급통계분석(16-17버전)", 2판, 지필출판사

LGM: More Advanced Topics (Cont'd)

- LGM에서 time-varying covariates에 해당하는 X_{ij} 변수 사용
  1. 시간에 따른 성장패턴이 linear trajectory가 아닐 수 있다.
  2. non-linearity and discontinuities in the growth trajectory
  3. Non-linear Growth Model을 설정하거나 Time-varying covariates을 포함할 수 있다.
  4. X_{ij}는 시간가변적인 변수일지라도 X_{ij} 변수가 y_{ij}에 미치는 효과는 시간불변적으로 가정한다 (constant across time periods).

LGM: Stata 실습

- Stata에서 LGM을 추정하는 방법
  1. 멀티레벨 모형 접근방법: `mixed` 명령어

     - Long-type 패널데이터 구조

  2. 구조방정식 모형 접근방법: `sem` 명령어

     - Wide-type 패널데이터 구조

- 본 강의에서는 멀티레벨 모형 접근만 설명

LGM: Stata 실습 (Cont'd)

- Frequency table : $Time_{ij}$

```
. tab time
```

time	Freq.	Percent	Cum.
1	333	13.30	13.30
2	384	15.34	28.65
3	395	15.78	44.43
4	379	15.14	59.57
5	358	14.30	73.87
6	346	13.82	87.69
7	308	12.31	100.00
Total	2,503	100.00	

LGM: Stata 실습 (Cont'd)

- Model 1: No random effects

```
. mixed lexp time , nolog mle
Mixed-effects ML regression                     Number of obs     =      2,460
                                                Wald chi2(1)      =     208.27
Log likelihood = -4239.3285                     Prob > chi2       =     0.0000
```

lexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
time	-.2033619	.0140914	-14.43	0.000	-.2309806	-.1757432
_cons	14.45033	.0615846	234.64	0.000	14.32963	14.57104

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
var(Residual)	1.838098	.0524102	1.738194	1.943745

- OLS로 추정결과와 일치

  Q13) *time* 변수 coefficient 해석?

LGM: Stata 실습 (Cont'd)

- Model 2: Random intercept

```
. mixed lexp time ||pid: , nolog mle
Mixed-effects ML regression                     Number of obs     =      2,460
Group variable: pid                             Number of groups  =        669
                                                Obs per group:
                                                              min =          1
                                                              avg =        3.7
                                                              max =          7
                                                Wald chi2(1)      =     192.16
Log likelihood = -4079.7761                     Prob > chi2       =     0.0000
```

lexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
time	-.1910689	.0137836	-13.86	0.000	-.2180843	-.1640536
_cons	14.44175	.0672178	214.85	0.000	14.31001	14.57349

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pid: Identity				
var(_cons)	.6471241	.0591044	.5410583	.7739825
var(Residual)	1.21973	.0409341	1.142082	1.302656

```
LR test vs. linear model: chibar2(01) = 319.10        Prob >= chibar2 = 0.0000
```

LGM: Stata 실습 (Cont'd)

- `xtreg, re mle` 로 추정한 결과와 일치
- y_{ij} 변수의 serial correlation을 가정한 것과 같다.

Q14) LR 가설검정 해석은?

LGM: Stata 실습 (Cont'd)

- Model 3: Random intercept and slope

```
. mixed lexp time ||pid:time , nolog mle cov(un)
Mixed-effects ML regression                     Number of obs     =      2,460
Group variable: pid                             Number of groups  =        669
                                                Obs per group:
                                                              min =          1
                                                              avg =        3.7
                                                              max =          7
                                                Wald chi2(1)      =     183.23
Log likelihood = -4068.3452                     Prob > chi2       =     0.0000
```

lexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
time	-.1946167	.0143773	-13.54	0.000	-.2227957	-.1664376
_cons	14.45059	.0623515	231.76	0.000	14.32838	14.57279

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pid: Unstructured				
var(time)	.0108782	.0069156	.0031291	.0378172
var(_cons)	.3802387	.1281438	.1964236	.7360695
cov(time,_cons)	.0147934	.0279521	-.0399918	.0695785
var(Residual)	1.178614	.0431777	1.096954	1.266353

```
LR test vs. linear model: chi2(3) = 341.97                 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

LGM: Stata 실습 (Cont'd)

- `cov(ind)` 옵션을 사용하면?
- $corr(u_{0j}, u_{1j})$ 추정

```
. estat recovariance, corr
Random-effects correlation matrix for level pid
```

	time	_cons
time	1	
_cons	.2300176	1

Q15) positive correation의 의미는?

LGM: Stata 실습 (Cont'd)

- Model 4: Curvilinear growth pattern

```
. mixed lexp time c.time#c.time ||pid:time c.time#c.time , nolog mle cov(ind)
Mixed-effects ML regression                     Number of obs     =      2,460
Group variable: pid                             Number of groups  =        669
                                                Obs per group:
                                                              min =          1
                                                              avg =        3.7
                                                              max =          7
                                                Wald chi2(2)      =     336.37
Log likelihood = -3996.8151                     Prob > chi2       =     0.0000
```

lexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
time	-.8276777	.0536407	-15.43	0.000	-.9328116	-.7225439
c.time#c.time	.0808117	.0066091	12.23	0.000	.067858	.0937653
_cons	15.37737	.0976187	157.52	0.000	15.18604	15.5687

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pid: Independent				
var(time)	.0159826	.0034648	.01045	.0244444
var(time#time)	1.37e-10	4.45e-08	4.9e-286	3.9e+265
var(_cons)	.4419512	.0644536	.332075	.5881831
var(Residual)	1.076242	.0374498	1.005289	1.152203

```
LR test vs. linear model: chi2(3) = 396.61                    Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

LGM: Stata 실습 (Cont'd)

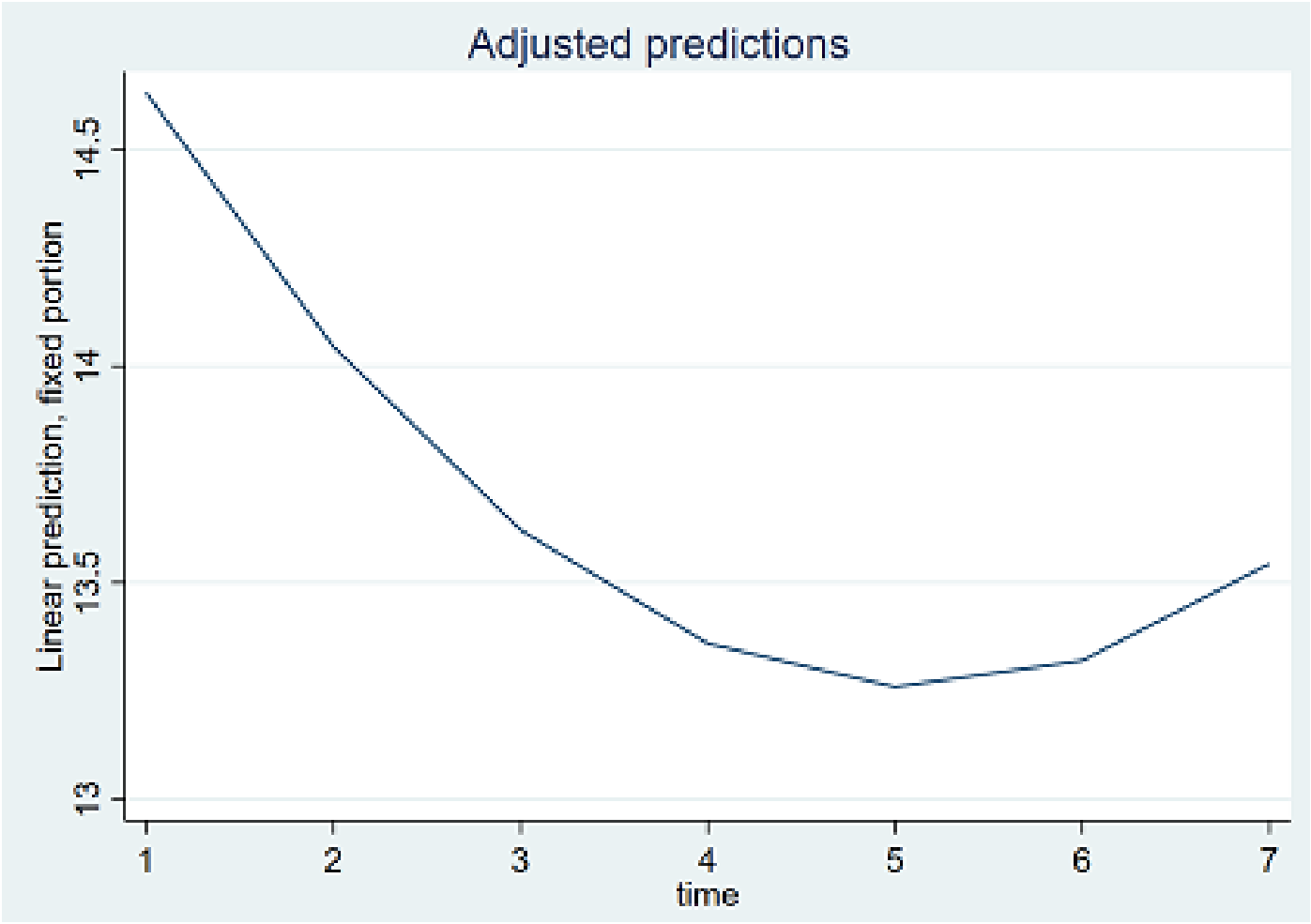
- *c.time#c.time* coefficient 해석은?
- 평균적 성장곡선 Prediction

```
. margins , at(time=(1(1)7)) atmeans noatlegend
Adjusted predictions                                  Number of obs = 2,460
Expression: Linear prediction, fixed portion, predict()
```

	Delta-method					
	Margin	std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	14.63051	.0597655	244.80	0.000	14.51337	14.74764
2	14.04526	.0433938	323.67	0.000	13.96021	14.13031
3	13.62164	.0439897	309.66	0.000	13.53542	13.70786
4	13.35965	.0474217	281.72	0.000	13.2667	13.45259
5	13.25927	.0489205	271.04	0.000	13.16339	13.35516
6	13.32052	.0546411	243.78	0.000	13.21343	13.42762
7	13.5434	.0754439	179.52	0.000	13.39553	13.69127

```
. marginsplot, noci recast(line)
Variables that uniquely identify margins: time
```

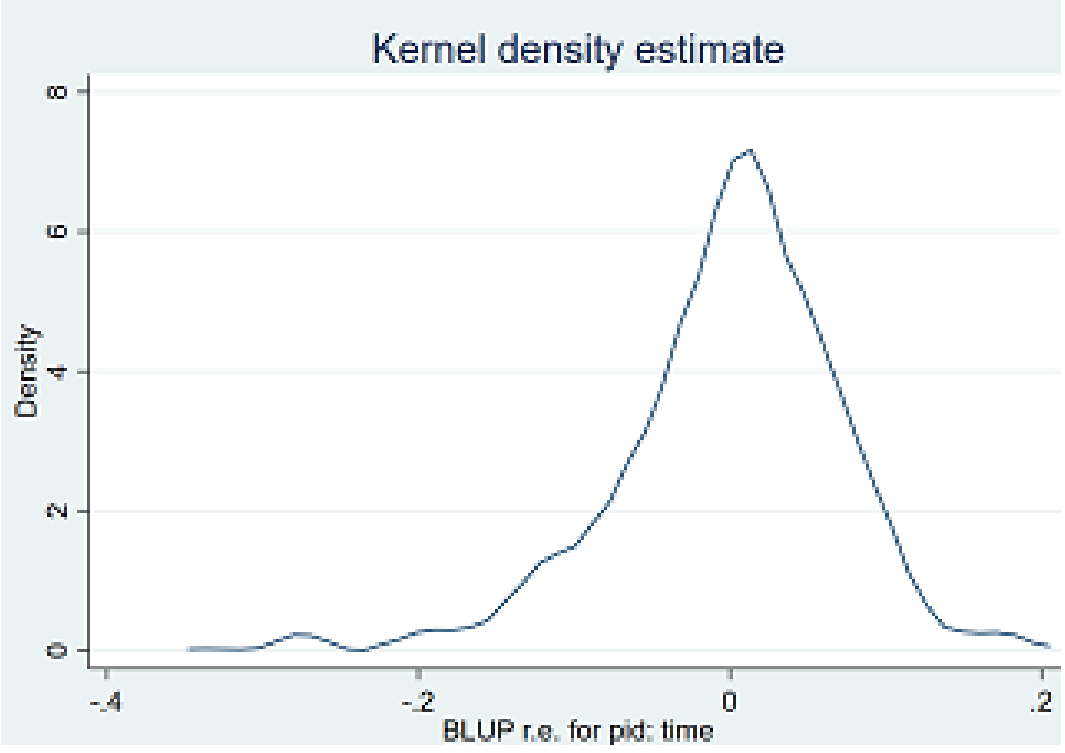
LGM: Stata 실습 (Cont'd)



LGM: Stata 실습 (Cont'd)

```
. predict re*, reffects
(7 missing values generated)
(7 missing values generated)
(7 missing values generated)
. describe re1 re2 re3
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
re1             float   %9.0g                 BLUP r.e. for pid: time
re2             float   %9.0g                 BLUP r.e. for pid: c.time#c.time
re3             float   %9.0g                 BLUP r.e. for pid: _cons
. kdensity re1
```

LGM: Stata 실습 (Cont'd)



Q16) *re*1 변수(random slope)의 분포 그래프에 대한 해석?

LGM: Stata 실습 (Cont'd)

- Model 6: Different time effects between Male and Female

```
. mixed lexp i.male time i.male#c.time ||pid: time , nolog mle cov(ind)
Mixed-effects ML regression                     Number of obs     =      2,460
Group variable: pid                             Number of groups  =        669
                                                Obs per group:
                                                              min =          1
                                                              avg =        3.7
                                                              max =          7
                                                Wald chi2(3)      =     181.59
Log likelihood =  -4067.938                     Prob > chi2       =     0.0000
```

lexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.male	.1290479	.1266032	1.02	0.308	-.1190897	.3771855
time	-.1861607	.0197968	-9.40	0.000	-.2249617	-.1473597
male#c.time						
1	-.0196339	.0292263	-0.67	0.502	-.0769164	.0376486
_cons	14.38859	.0877493	163.97	0.000	14.21661	14.56058

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pid: Independent				
var(time)	.0140692	.0033998	.0087614	.0225924
var(_cons)	.4407346	.065634	.3291672	.5901164
var(Residual)	1.170433	.0405914	1.093519	1.252757

```
LR test vs. linear model: chi2(2) = 342.56                    Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

LGM: Stata 실습 (Cont'd)

- *male#c.titme* coefficient의 해석은?
- 상위레벨 변수(성별)가 growth pattern에 미치는 조절효과로 이해할 수 있다.

LGM: Stata 실습 (Cont'd)

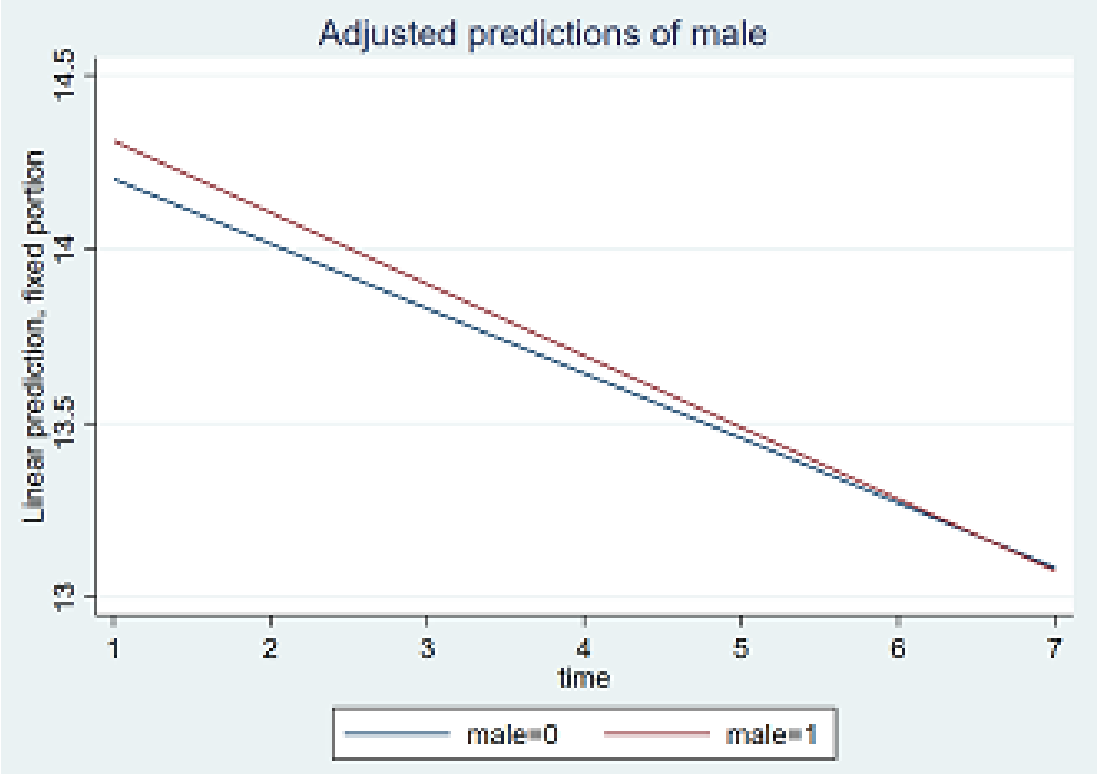
- 평균적 성장곡선 Prediction by sex

```
. margins male, at(time=(1(1)7)) atmeans noatlegend
Adjusted predictions                                  Number of obs = 2,460
Expression: Linear prediction, fixed portion, predict()
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at#male						
1 0	14.20243	.0731228	194.23	0.000	14.05911	14.34575
1 1	14.31185	.0756246	189.25	0.000	14.16362	14.46007
2 0	14.01627	.0614636	228.04	0.000	13.8958	14.13674
2 1	14.10605	.0635158	222.09	0.000	13.98156	14.23054
3 0	13.83011	.0547032	252.82	0.000	13.72289	13.93733
3 1	13.90026	.0572186	242.93	0.000	13.78811	14.0124
4 0	13.64395	.0546894	249.48	0.000	13.53676	13.75114
4 1	13.69446	.0586362	233.55	0.000	13.57954	13.80939
5 0	13.45779	.0614267	219.09	0.000	13.33739	13.57818
5 1	13.48867	.0672828	200.48	0.000	13.3568	13.62054
6 0	13.27163	.0730711	181.63	0.000	13.12841	13.41484
6 1	13.28287	.0808719	164.25	0.000	13.12437	13.44138
7 0	13.08547	.0876889	149.23	0.000	12.9136	13.25733
7 1	13.07708	.0973555	134.32	0.000	12.88626	13.26789

```
. marginsplot, noci recast(line)
Variables that uniquely identify margins: time male
```

LGM: Stata 실습 (Cont'd)



Q16) 위 prediction 그래프에 대한 해석은?

LGM: Stata 실습 (Cont'd)

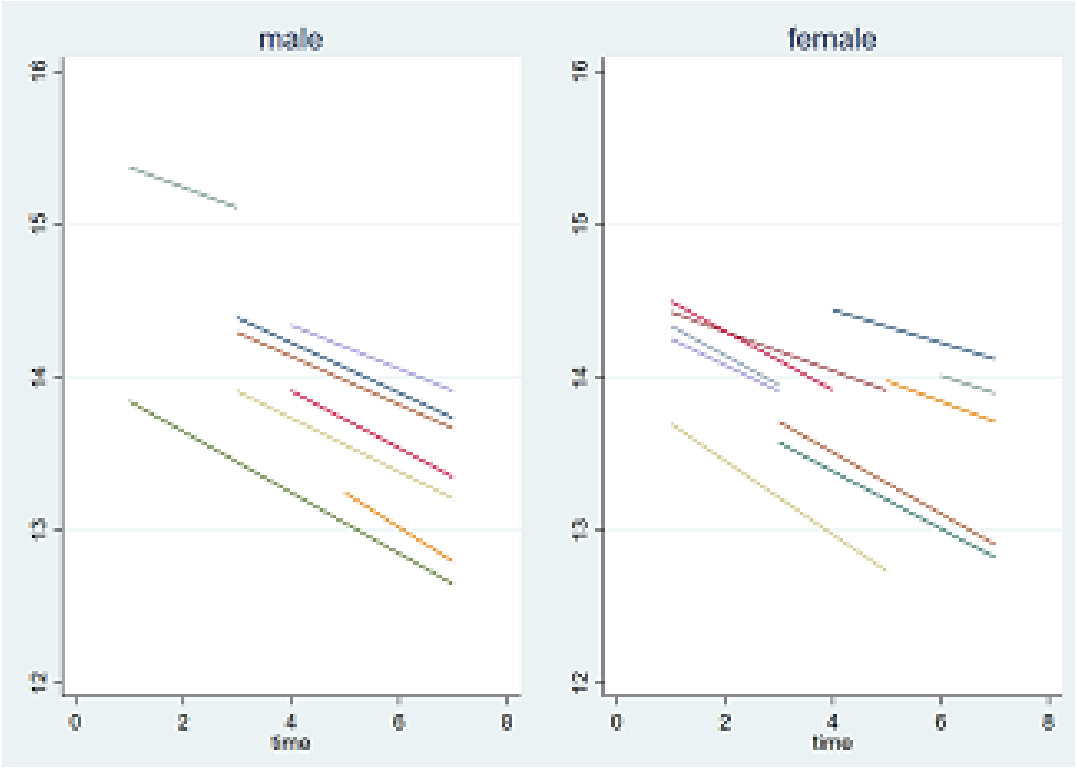
- pid=1 ∼ 20번 표본에 대해 추정된 individual trajectory를 그래프로 작성

```
. qui mixed lexp i.male time i.male#c.time ||pid: time , nolog mle cov(ind)
. predict yhat, fitted
(7 missing values generated)
. xtline yhat if pid<=20 & male==1, overlay legend(off) ytitle("") name(graph1, replace) title(male)
. xtline yhat if pid<=20 & male==0, overlay legend(off) ytitle("") name(graph2, replace) title(female)
. graph combine graph1 graph2, ycommon
```

- `predict` 명령문에서 `fitted` 옵션을 사용하면

$$\widehat{y}_{ij} = \underbrace{\widehat{\gamma}_0 + \widehat{\delta}_0 Male_j + \widehat{\gamma}_1 Time_{ij} + \widehat{\delta}_1 (Male_j \times Time_{ij})}_{\text{Fixed part}} + \underbrace{\widehat{u}_{0j} + \widehat{u}_{1j} Time_{ij}}_{\text{Random part}}$$

LGM: Stata 실습 (Cont'd)



참석해 주셔서 감사드립니다

Thank You !