Multilevel Data Modeling

2022 KHP 자료설명회 특강

July 22, 2022

Table of Contents

smart_khp 패키지

- 1기 의료패널 데이터: 2008~2018년(11개년) 패널데이터

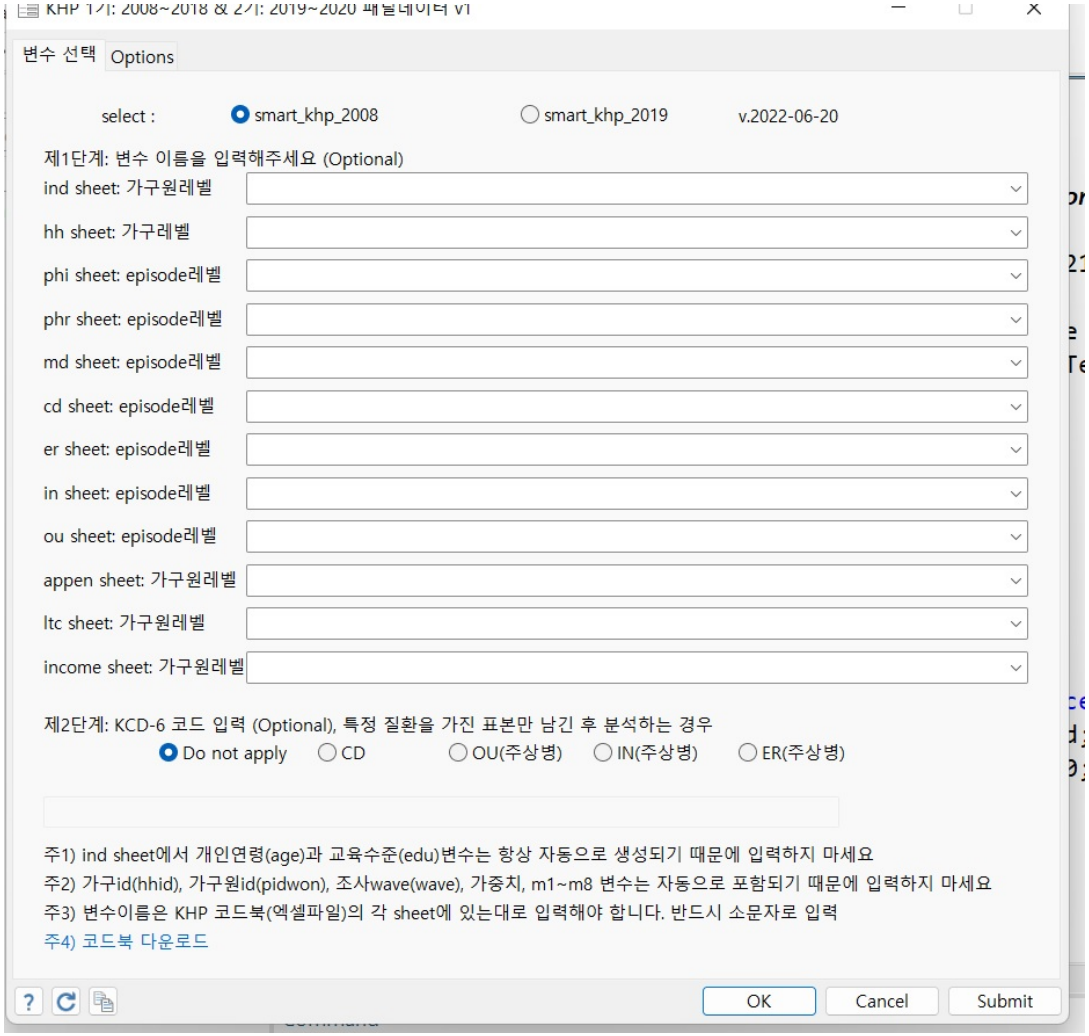
  2기 의료패널 데이터: 2019~2020년 (2개년) 패널데이터

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

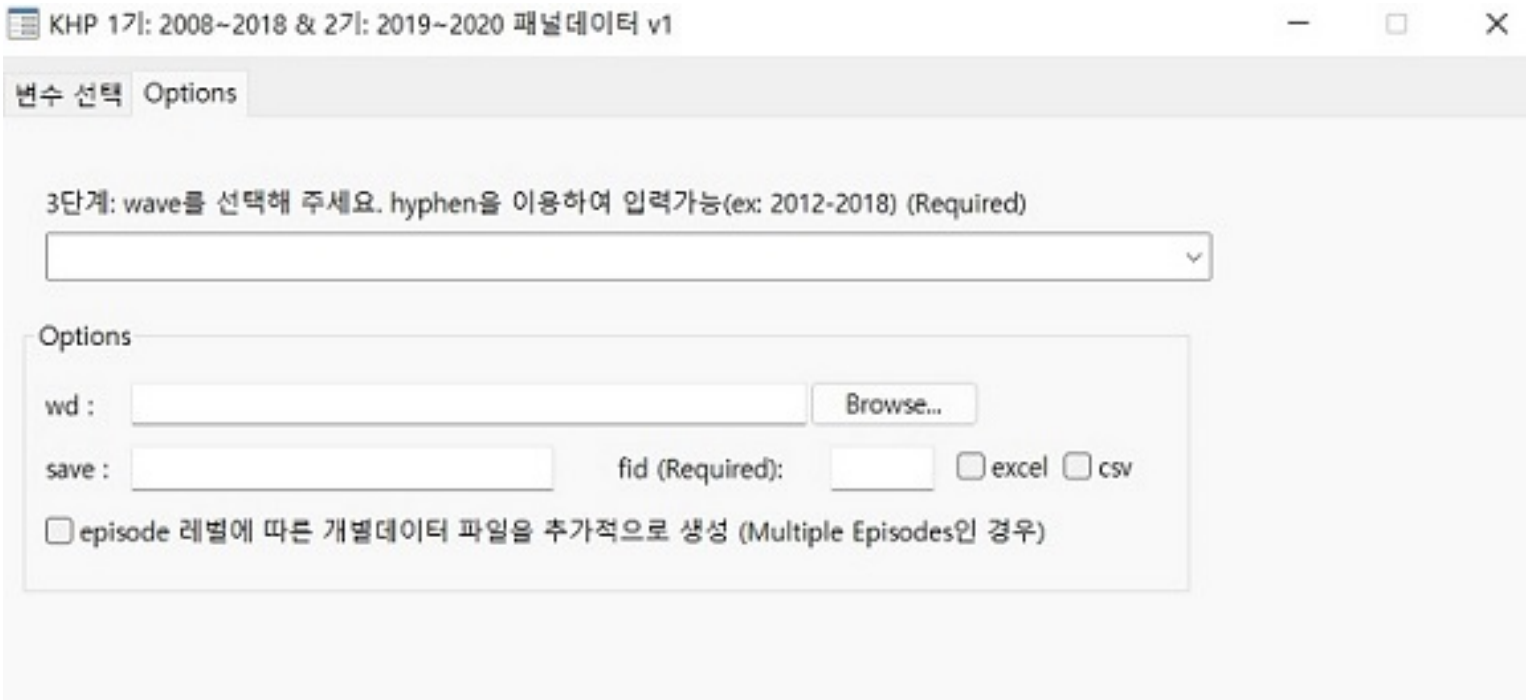
  2기 KHP: 4개 부문 ($4 \times 2 = 8$개 data files)

- KHP 1기 & 2기 데이터를 손쉽게 패널데이터로 만들 수 있는 Stata 명령어 패키지: smart_khp

smart_khp 패키지 (Cont'd)



smart_khp 패키지 (Cont'd)



smart_khp 패키지 (Cont'd)

- 1기 KHP 데이터: smart_khp_2008 명령어
- 2기 KHP 데이터: smart_khp_2019 명령어
- 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_2008 , ind(c3) appen(s2 s17) ///
ou_kcd kcd(I10 K30) wave(2012-2018) fid(비밀번호)
```

note) `fid( )` 옵션 required

smart_khp 패키지 (Cont'd)

- 2022년 12월 중 Stata 코딩파일과 매뉴얼 공개 예정
  : 추후 KHP 홈페이지 공지

- 민인식(2021), 한국의료패널(KHP) 활용을 위한 Stata 패키지 개발, *의료경영학 연구*, 15(1), 25-35

Introduction to Multilevel Data

- 멀티레벨 데이터는 횡단면 시점에서 nested data structure와 시계열 시점에서 panel data로 구분
- 횡단면 데이터: 어느 한 시점에서 상위레벨 그룹(cluster) 내에서 개인(individual)에 대한 random sample이다.

  - 상위레벨 그룹: schools. hospitals, states, areas 등

Introduction to Multilevel Data (Cont'd)

- 패널데이터(종단면 데이터): 하나의 개인(가구)을 여러 시점에서 수집한 데이터

  - 상위레벨: 개인, 하위레벨: time points

- Two-level data 구조 뿐 아니라 Three-level data 구조도 가능

  - level 3: 개인, level 2: time, level 1: 병원방문 episode

  - level 3: 병원, level 2: 개인, level 1: time

  - level 3: 가구, level 2: 가구원, level 1: time

Introduction to Multilevel Data (Cont'd)

- 3-level data 예시 : 가구원-시점-병원방문

pidwon	year	epi_type	IND_age	IND_i_medi~1	OU_ou29_2	OU_ou29_4
1000102	2015	OU	76	822100	4500	10710
1000102	2015	OU	76	822100	1500	10220
1000102	2016	OU	77	933400	5300	12620
1000102	2016	OU	77	933400	5200	12240

```
smart_khp_2008  , ind( i_medicalexp1) ///
ou( ou29_2 ou29_4) wave(2015-2018)
```

Introduction to Multilevel Data (Cont'd)

- 멀티레벨 데이터의 특징

  - 상위레벨(학교 또는 가구) 내에 속한 하위레벨 (학생 또는 가구원)은 서로 상관관계를 가지고 있다.

  - 이러한 상관관계를 고려하여 모수를 추정해야 하고 standard error를 얻어야 할 필요가 있다.

- Pooled/Separate Regression과 차이

  - Pooled Regression: 상위레벨 개체(entity) 간 서로 variations이 있다는 것을 무시한다.

  - Separate regression: sample size problem and lack of generalization

Introduction to Multilevel Data (Cont'd)

- Multilevel Regression model의 다른 이름

  - Mixed Linear Model

  - Mixed Effects Model

  - Hierarchical Regression Model

  - Panel Regression Model : 하위레벨(level 1)로 time(시간)을 선택하는 경우

  Q1) 횡단면 멀티레벨과 시계열 멀티레벨(패널 데이터)의 차이점은?

Random Intercept Model

- varying-intercept 모형이라고도 부른다.

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

  where $\alpha_j = \alpha_0 + u_j$

- u_j: level 2(개인) 이질성에 해당하는 오차항(level-2 errors)

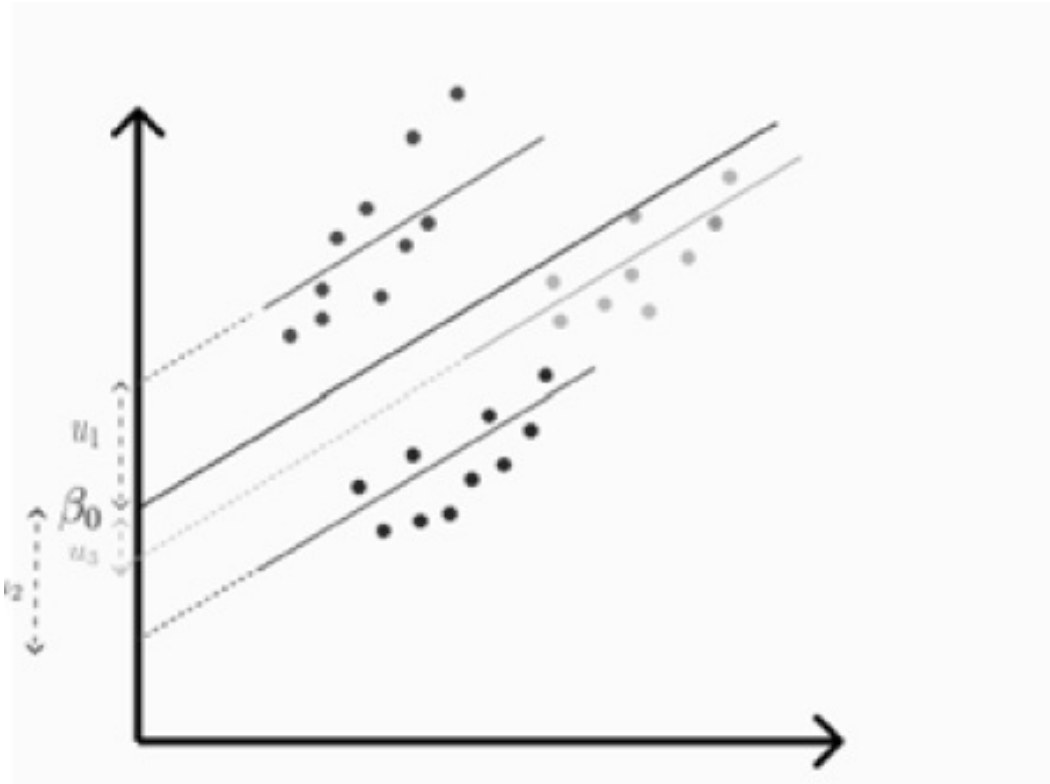
$$u_j \sim N(0, \sigma_u^2)$$

Random Intercept Model (Cont'd)

- ϵ_{ij}: level 1이 시간(time)이라면 시간가변적인 오차항(level-1 error)

$$\epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2)$$

Random Intercept Model (Cont'd)



- 상위개체별로 서로 다른 상수항을 가지는 것을 허용하며 추정의 목적은 random intercept term에 해당하는 α_j의 분산인 σ_u^2을 추정하는 것이다.

Random Intercept Model (Cont'd)

- Stata 14버전 이후에서는 **mixed** 명령어를 이용하여 멀티레벨&선형회귀모형인 경우 추정결과를 얻는다.

```
. mixed lmexp CD_cdnum || pidwon:, nolog mle
Mixed-effects ML regression                     Number of obs     =     28,055
Group variable: pidwon                          Number of groups  =      5,609
                                                Obs per group:
                                                              min =          1
                                                              avg =        5.0
                                                              max =          7
                                                Wald chi2(1)      =     695.40
Log likelihood = -65223.187                     Prob > chi2       =     0.0000
```

lmexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
CD_cdnum	.2559294	.0097052	26.37	0.000	.2369076	.2749512
_cons	11.60487	.0467794	248.08	0.000	11.51319	11.69656

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pidwon: Identity				
var(_cons)	3.916681	.0989977	3.727376	4.115599
var(Residual)	4.434533	.0422045	4.35258	4.518029

LR test vs. linear model: chibar2(01) = 7461.27 Prob >= chibar2 = 0.0000

Random Intercept Model (Cont'd)

- Intra-class correlation

$$corr(y_{ij}, y_{kj}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$$

```
. estat icc
Residual intraclass correlation
```

Level	ICC	Std. err.	[95% conf.	interval]
pidwon	.4689954	.0070269	.4552501	.4827879

Q2) ICC 결과 해석:

note) 패널데이터인 경우에는 mixed 대신 xtreg를 사용해도 Random Intercept 모형을 추정할 수 있다.

Random Coefficient Model

- varying-coefficient 모형이라고도 부른다.
  - random intercept와 random slope가 모두 포함된 모형이다.

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$$

where

$$\alpha_j = \alpha_0 + u_{0j}$$

$$\beta_j = \beta_1 + u_{1j}$$

$$\Longrightarrow y_{ij} = \underbrace{\alpha_0 + \beta_1 x_{ij}}_{\text{Fixed Part}} + \underbrace{u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}}_{\text{Random Part}}$$

Random Coefficient Model (Cont'd)

- 상수항(intercept term) 뿐만 아니라 x_{ij} 변수의 기울기도 상위레벨 그룹에 따라 다르다는 것을 허용한다.

  - x_{ij} 설명변수가 각 상위레벨 그룹에서 서로 다른 한계효과(marginal effects)를 가질 수 있다.

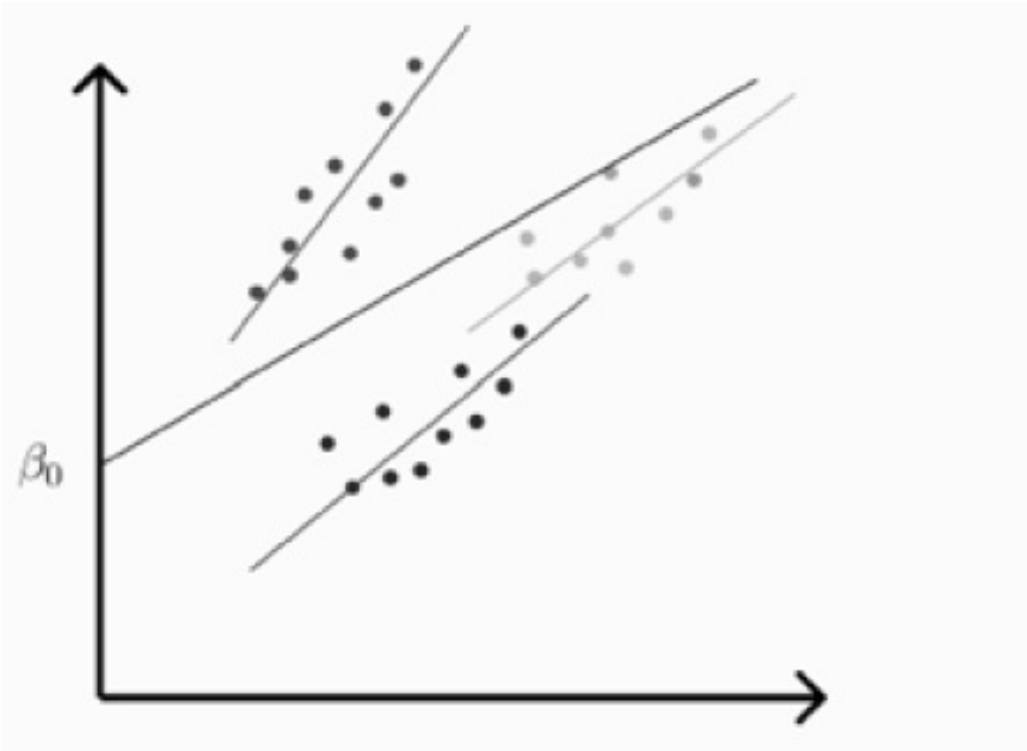
- u_{0j}, u_{1j}: level 2(개인) 이질성에 해당하는 오차항(level-2 errors)

$$u_{0j} \sim N(0, \sigma^2_{u0}), \quad u_{1j} \sim N(0, \sigma^2_{u1})$$

same level covariance:

$$cov(u_{0j}, u_{1j}) = 0 \textbf{ OR } cov(u_{0j}, u_{1j}) \neq 0$$

Random Coefficient Model (Cont'd)



Random Coefficient Model (Cont'd)

```
. mixed lmexp exercise|| pidwon: exercise  , nolog mle cov(ind)
Mixed-effects ML regression                     Number of obs     =     27,418
Group variable: pidwon                          Number of groups  =      5,510
                                                Obs per group:
                                                              min =          1
                                                              avg =        5.0
                                                              max =          7
                                                Wald chi2(1)      =       7.15
Log likelihood = -62400.811                     Prob > chi2       =     0.0075
```

lmexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
exercise	-.029814	.0111507	-2.67	0.008	-.0516691	-.0079589
_cons	12.65324	.0305905	413.63	0.000	12.59329	12.7132

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pidwon: Independent				
var(exercise)	.0109698	.0057563	.0039223	.0306801
var(_cons)	4.050218	.0985007	3.861689	4.247951
var(Residual)	3.904069	.0382299	3.829854	3.979722

```
LR test vs. linear model: chi2(2) = 8942.85                   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

Random Coefficient Model (Cont'd)

Q3) same level covariance $\neq 0$으로 가정하면 어떤 옵션을 사용하는가?

Q4) Random slope만 포함한 모형을 추정하고자 하는 경우는?

- Intra-class correlation

$$corr(y_{ij}, y_{kj}) = \frac{\sigma^2_{u0} + x_{ij}x_{kj}\sigma^2_{u1}}{\sigma^2_{u0} + x_{ij}x_{kj}\sigma^2_{u1} + \sigma^2_{\epsilon}}$$

```
. estat icc
Conditional intraclass correlation
```

Level	ICC	Std. err.	[95% conf.	interval]
pidwon	.5091868	.0067914	.4958726	.522488

```
Note: ICC is conditional on zero values of random-effects covariates.
```

Random Coefficient Model (Cont'd)

note) 멀티레벨 데이터가 패널데이터인 경우에 mixed 대신 xtreg를 사용해서 random coefficient 모형을 추정할 수 없다.

Three-level Model

- level-3: 가구(h), level-2: 가구원(j), level-1: time(i) 으로 설정하자.

  - 논의의 편의상 random intercept 모형 예시

$$y_{ijh} = \alpha_{jh} + \beta x_{ijh} + \epsilon_{ijh}$$

  where

$$\alpha_{jh} = \alpha_0 + u_{0jh} + u_{1h}$$

- intercept term은 최상위레벨인 가구(hhid)에 따라 달라질 수 있으며 가구 내에서도 중간레벨인 가구원(pidwon)에 따라 변할 수 있다.

Three-level Model (Cont'd)

- u_{0jh} : level 2(가구원) 이질성을 포함하는 오차항 (level-2 error)

  u_{1h}: level 3(가구) 이질성에 해당하는 오차항(level-3 errors)

- cross-level covariance=0으로 가정한다.

$$cov(u_{0jh}, u_{1h}) = 0, cov(u_{0jh}, \epsilon_{ijh}) = 0, cov(u_{1h}, \epsilon_{ijh}) = 0$$

Three-level Model (Cont'd)

```
. mixed lmexp exercise|| hhid:|| pidwon:  , nolog mle cov(ind)
note: single-variable random-effects specification in pidwon equation; covariance
      structure set to identity.
Mixed-effects ML regression                     Number of obs     =     27,418
        Grouping information
```

Group variable	No. of groups	Observations per group Minimum	Average	Maximum
hhid	3,741	1	7.3	21
pidwon	5,552	1	4.9	7

```
                                                Wald chi2(1)      =       7.31
Log likelihood = -62386.065                     Prob > chi2       =     0.0068
```

lmexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
exercise	-.028782	.0106429	-2.70	0.007	-.0496418	-.0079223
_cons	12.64502	.0319732	395.49	0.000	12.58236	12.70769

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
hhid: Identity				
var(_cons)	.7811522	.12201	.5751544	1.06093
pidwon: Identity				
var(_cons)	3.272652	.136123	3.016441	3.550625
var(Residual)	3.917438	.0375949	3.844442	3.99182

```
LR test vs. linear model: chi2(2) = 8972.35                   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

Three-level Model (Cont'd)

```
. estat icc
Residual intraclass correlation
```

Level	ICC	Std. err.	[95% conf.	interval]
hhid	.0979963	.015047	.0722031	.1316957
pidwon|hhid	.5085536	.0068039	.4952154	.5218797

- Intra-class correlation

$$corr(y_{ijh}, y_{kjh}) = \frac{\sigma^2_{u0} + \sigma^2_{u1}}{\sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_{\epsilon}}$$

$$corr(y_{ijh}, y_{kfh}) = \frac{\sigma^2_{u1}}{\sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_{\epsilon}}$$

Cross-level Interaction Model

- 상위레벨(가구원), 하위레벨(time)으로 가정하자.
- 상위레벨 covariate: 성별 변수(z_j)는 모형에 어떻게 포함될 수 있는가?

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$$

where

$$\alpha_j = \alpha_0 + \lambda_0 z_j + u_{0j}$$

$$\beta_j = \beta_1 + \lambda_1 z_j + u_{1j}$$

$$\Longrightarrow y_{ij} = \alpha_0 + \lambda_0 z_j + \beta_1 x_{ij} + \lambda_1 (x_{ij} \times z_j) + u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}$$

Cross-level Interaction Model (Cont'd)

- $(x_{ij} \times z_j)$항이 cross-level interaction term이 된다.
- 이러한 모형을 통해 상위레벨(z_j) 변수의 조절효과(moderation effect)를 판단할 수 있다.

  - 하위레벨 변수 x_{ij}가 종속변수 y_{ij}에 미치는 효과는 상위레벨 변수 z_j에 의해 영향을 받는가?

  Q5) 상위레벨 변수 z_j 변수만 포함하는 모형(interaction term이 없는)은 어떻게 이해할 수 있는가?

Cross-level Interaction Model (Cont'd)

```
. mixed lmexp i.female exercise i.female#c.exercise ||pidwon: exercise  , nolog mle
Mixed-effects ML regression                     Number of obs     =     27,418
Group variable: pidwon                          Number of groups  =      5,510
                                                Obs per group:
                                                              min =          1
                                                              avg =        5.0
                                                              max =          7
                                                Wald chi2(3)      =      38.95
Log likelihood = -62384.977                     Prob > chi2       =     0.0000
```

lmexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.female	.3209682	.0616303	5.21	0.000	.200175	.4417615
exercise	-.0385731	.0140788	-2.74	0.006	-.066167	-.0109791
female#c.exercise						
1	.0334481	.0230735	1.45	0.147	-.0117751	.0786713
_cons	12.47015	.046603	267.58	0.000	12.37881	12.56149

Random-effects parameters	Estimate	Std. err.	[95% conf.	interval]
pidwon: Independent				
var(exercise)	.0106379	.0057313	.0037005	.0305807
var(_cons)	4.018743	.0979203	3.831333	4.21532
var(Residual)	3.904994	.0382435	3.830753	3.980674

```
LR test vs. linear model: chi2(2) = 8861.59                   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

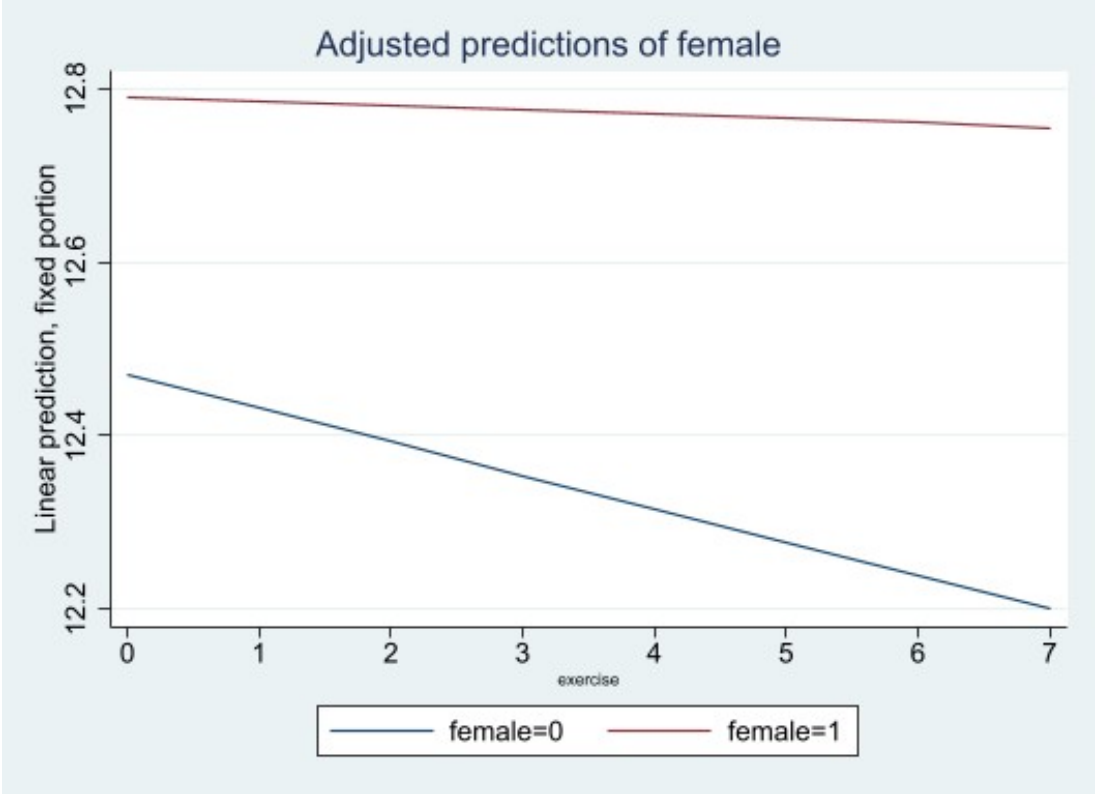
Cross-level Interaction Model (Cont'd)

Q6) 성별변수의 조절효과에 대한 유의성은?

- Prediction Graph

```
margins female, at(exercise=(0(1)7)) ///
atmeans noatlegend

marginsplot, noci recast(line) ///
xtitle("exercise", size(vsmall))
```

Cross-level Interaction Model (Cont'd)



note) margins 명령문에서는 fixed part에 대한 예측치만 이용하여 그래프를 작성한다. 즉 random effects=0으로 간주한다.

Multilevel Logit Model

- Discrete and Limited Dependent Variables

  - Binary variable
  - Count variable
  - Ordinal variable
  - Unordered variable
  - Censored variable

- Stata 17버전에서는 횡단면 모형 뿐 아니라 멀티레벨(또는 패널) 구조 데이터에서도 위 종속변수 모형을 추정할 수 있다.

Multilevel Logit Model (Cont'd)

- 종속변수가 binary variable인 경우 : latent response model

$$y_{ij}^{*} = \alpha + \beta x_{ij} + u_j + \epsilon_{ij}$$

where

$$y_{ij} = \begin{cases} 1, & \text{if } y_{ij}^{*} > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

- level-2(상위레벨) error(heterogeneity)를 포함하여 within-group correlation을 고려할 수 있다. $\Longrightarrow$ random effects 모형

$$u_j \sim N(0, \sigma_u^2)$$

Multilevel Logit Model (Cont'd)

- level-1 오차항 ϵ_{ij} 분포를 표준정규분포 $\sim N(0,1)$
  $\Longrightarrow$ Multilevel Probit 모형
- level-1 오차항 ϵ_{ij} 분포를 로지스틱분포 $\sim$ logistic$(0,\ \pi^2/3)$
  $\Longrightarrow$ Multilevel Logit 모형

$$F(\epsilon_{ij}) = \frac{\exp(\epsilon_{ij})}{1+\exp(\epsilon_{ij})} \quad \textbf{: logitstic CDF}$$

Multilevel Logit Model (Cont'd)

- 멀티레벨 로짓모형 하에서 $\Pr(y_{ij} = 1)$은 다음과 같이 쓸 수 있다.

$$\Pr_{ij} = \Pr(y_{ij} = 1 | x_{ij}, u_j) = F(\alpha + \beta x_{ij} + u_j)$$

where $F()$은 로지스틱 분포의 CDF이다.

- 승산(odds)

$$odds = \frac{\Pr_{ij}}{1 - \Pr_{ij}}$$

Multilevel Logit Model (Cont'd)

- 승산비(odds ratio): x_{ij} 변수가 1단위 증가할 때 승산에 미치는 영향

$$\textbf{odds ratio} = \exp(\beta)$$

Q6) conditional probability $\Pr(y_{ij}|x_{ij}, u_j)$와 unconditional probability $\Pr(y_{ij}|x_{ij})$은 어떻게 다른가?

Multilevel Logit Model (Cont'd)

- smart_khp 명령어를 이용하여 two-level(가구원-시간) 데이터를 만든다.

  1년 동안 상급종합병원 방문여부(s_hospital 변수)을 종속변수로 설정

```
smart_khp_2008, ind(i_medicalexp1 c3) hh(tot_h) ///
ou(ou11) wave(2012-2018)  save(khp_4)
```

Multilevel Logit Model (Cont'd)

- Random intercept 모형 추정

```
. xtset pidwon year
Panel variable: pidwon (unbalanced)
 Time variable: year, 2012 to 2018, but with gaps
         Delta: 1 unit
. tab s_hospital
```

s_hospital	Freq.	Percent	Cum.
0	105,540	88.19	88.19
1	14,136	11.81	100.00
Total	119,676	100.00	

```
. tab year s_hospital, row
```

Key
frequency *row percentage*

year	s_hospital 0	1	Total
2012	14,117 88.94	1,755 11.06	15,872 100.00
2013	13,238 89.21	1,601 10.79	14,839 100.00
2014	17,053 88.73	2,166 11.27	19,219 100.00
2015	16,057 88.57	2,073 11.43	18,130 100.00
2016	15,270 87.64	2,154 12.36	17,424 100.00
2017	14,995 87.26	2,189 12.74	17,184 100.00
2018	14,810 87.08	2,198 12.92	17,008 100.00
Total	105,540 88.19	14,136 11.81	119,676 100.00

Multilevel Logit Model (Cont'd)

```
. qui melogit s_hos i.female IND_age lincome, nolog
. melogit s_hos i.female IND_age lincome ||pidwon:, nolog
Mixed-effects logistic regression               Number of obs     =    108,079
Group variable: pidwon                          Number of groups  =     21,307
                                                Obs per group:
                                                              min =          1
                                                              avg =        5.1
                                                              max =          7
Integration method: mvaghermite                 Integration pts.  =          7
                                                Wald chi2(3)      =     927.63
Log likelihood = -26530.242                     Prob > chi2       =     0.0000
```

s_hospital	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.female	.1089579	.0574042	1.90	0.058	-.0035522	.2214681
IND_age	.0428379	.0014834	28.88	0.000	.0399305	.0457452
lincome	.0334365	.0249761	1.34	0.181	-.0155158	.0823888
_cons	-6.819633	.2462667	-27.69	0.000	-7.302307	-6.336959
pidwon						
var(_cons)	12.11447	.4339726			11.29307	12.99561

```
LR test vs. logistic model: chibar2(01) = 17012.85    Prob >= chibar2 = 0.0000
```

Q7) 맨 아래에 있는 LR test에 대한 해석은?

Multilevel Logit Model (Cont'd)

Q8) 위 추정결과에서 $var(\epsilon_{ij})$ 추정치가 없는 이유는?

- 승산비(odds ratio) 추정결과를 얻기 위해서는 `melogit, or`
- 같은 가구원 내 시점 간 종속변수의 dependency

$$corr(y_{ij}^*, y_{kj}^*) = \frac{\sigma_u^2}{(\pi^2/3) + \sigma_u^2}$$

```
. estat icc
Residual intraclass correlation
```

Level	ICC	Std. err.	[95% conf.	interval]
pidwon	.7864324	.0060166	.774403	.7979877

Multilevel Logit Model (Cont'd)

- Prediction

  1) fixed effects + the posterior mean of random effects

$$\Pr(y_{ij} = 1) = \frac{\exp(x_{ij}\beta + \widehat{u}_j)}{1 + \exp(x_{ij}\beta + \widehat{u}_j)}$$

  2) only fixed effects + the random effects(=0, prior mean)

$$\Pr(y_{ij} = 1) = \frac{\exp(x_{ij}\beta)}{1 + \exp(x_{ij}\beta)}$$

Multilevel Logit Model (Cont'd)

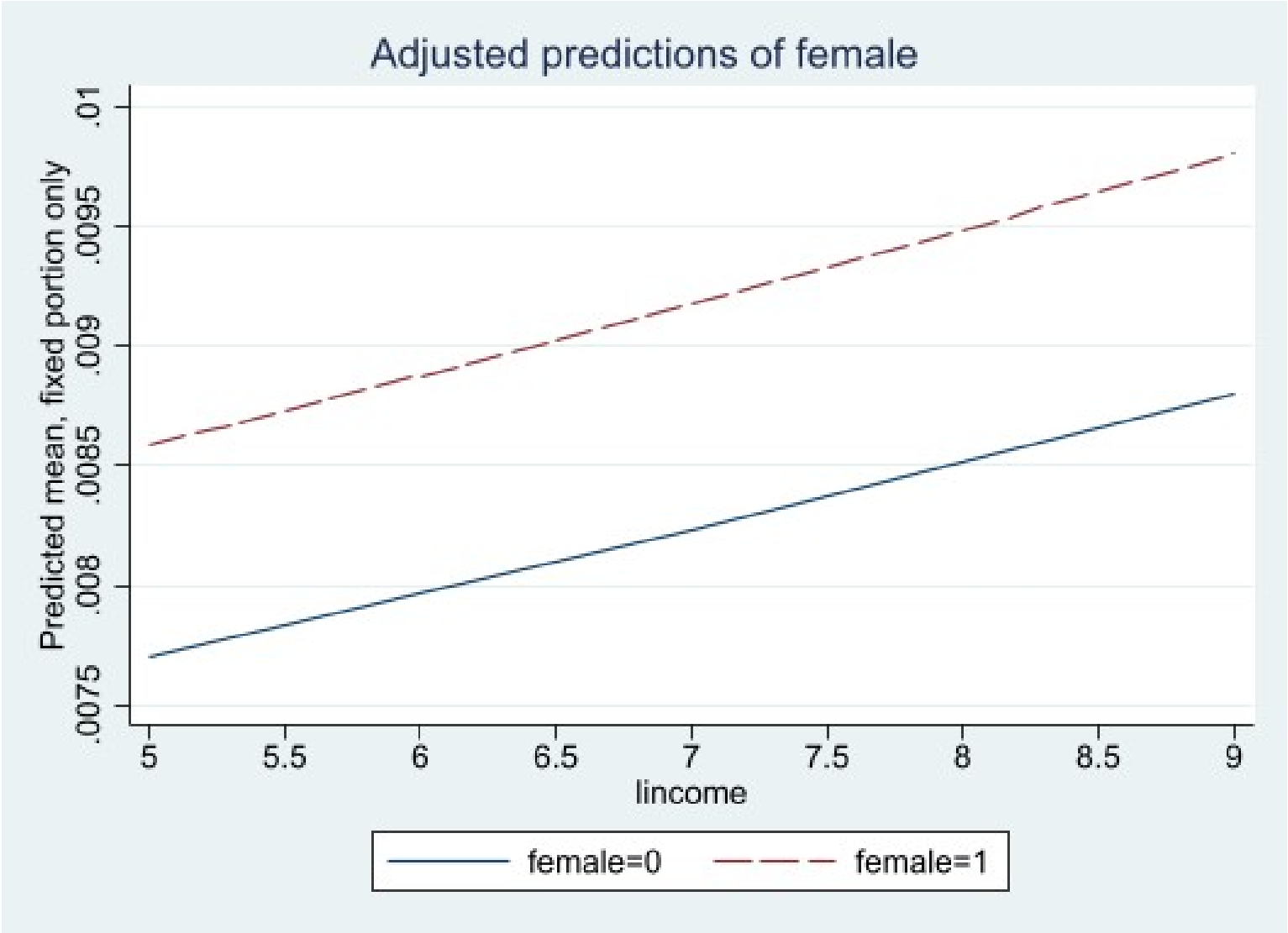
```
predict re*, reffects

predict pr1, mu

predict pr2, conditional(fixed)
```

Multilevel Logit Model (Cont'd)

- Prediction graph : only fixed effects

```
margins female, at(lincome=(5(0.5)9)) ///
atmeans noatlegend predict(conditional(fixed))

marginsplot, noci recast(line) ///
plot2opts(lpattern(longdash))
```

Multilevel Logit Model (Cont'd)

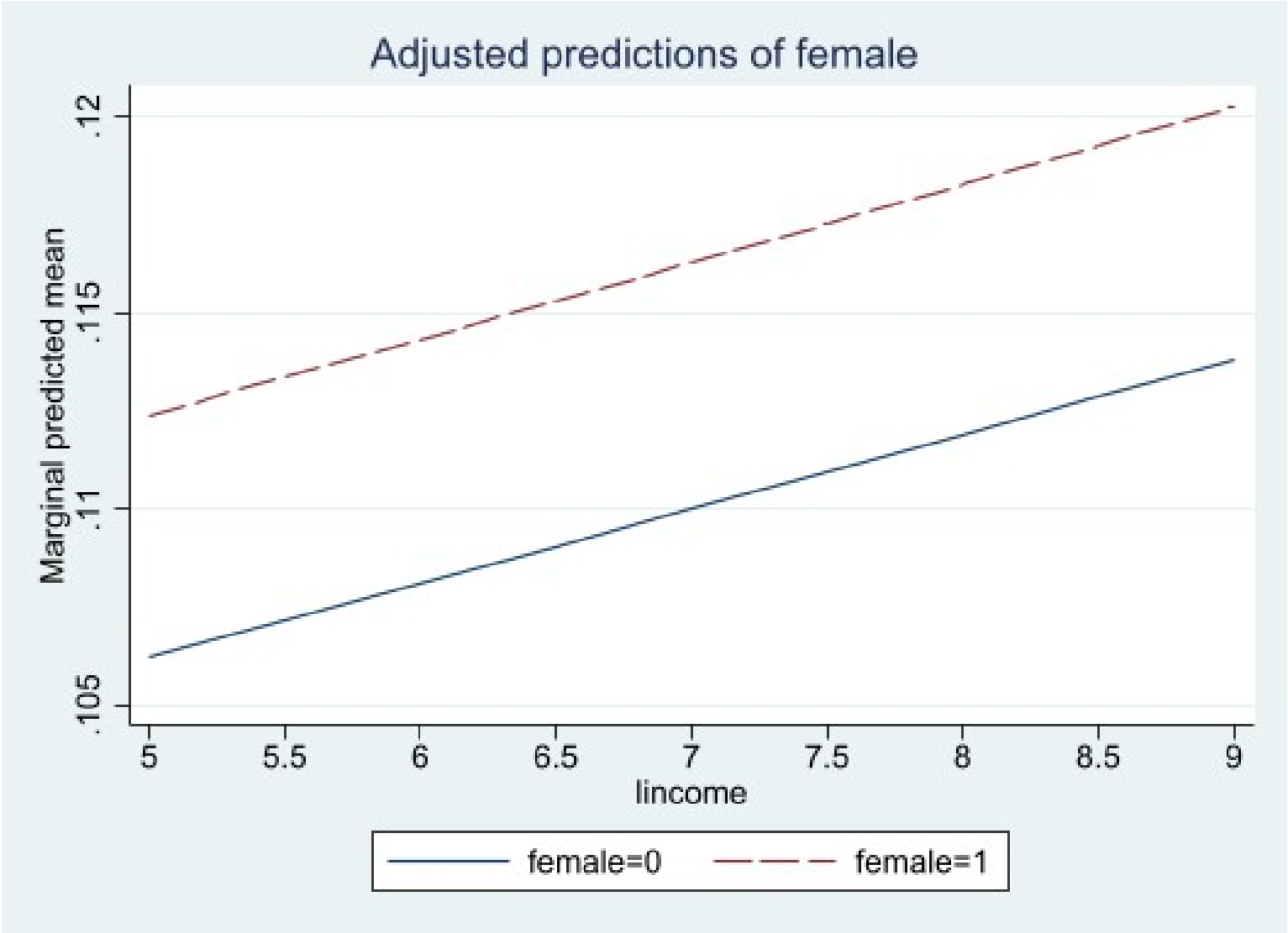


Multilevel Logit Model (Cont'd)

- Prediction graph : fixed effects + individual-specific heterogeneity

```
margins female, at(lincome=(5(0.5)9)) ///
atmeans noatlegend predict(mu)

marginsplot, noci recast(line) ///
plot2opts(lpattern(longdash))
```

Multilevel Logit Model (Cont'd)



Multilevel Logit Model (Cont'd)

- individual-specific heterogeneity(질환의 경중 등)을 포함하면 상급종합병원 방문 확률이 크게 높아진다는 것을 확인할 수 있다.

Multilevel Tobit Model

- 종속변수가 censored variable(절단형 변수)인 경우에 Tobit 모형을 사용한다.종속변수가 일정한 범위 내에서만 관찰된다.
- 특히 종속변수 값이 0으로 관찰되는 값이 많은 경우에 사용한다. 0보다 작은 값은 관찰되지 않고 모두 0으로 관찰된다. 따라서 0의 값을 갖더라도 모두 같은 0이라고 해석할 수 없다.
- censored observations을 제외하고 추정한다면 entire population에 대한 inference를 얻을 수 없는 문제가 발생한다.

Multilevel Tobit Model (Cont'd)

- 멀티레벨 데이터 구조 + 종속변수 censored인 경우 $\Longrightarrow$ 멀티레벨 토빗모형
- 패널데이터 구조(two-level 구조)인 경우에도 패널 토빗 또는 멀티레벨 토빗모형을 활용할 수 있다.
- Stata 명령어: metobit 또는 xttobit 명령어

  - 다만 xttobit 명령어에서는 random intercept 모형만 추정 가능하다

Multilevel Tobit Model (Cont'd)

- Latent Response Model Approach : Random-coefficient 모형

$$y_{ij}^{*} = \alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}$$

where observed dependent variable y_{ij}

$$y_{ij} = \begin{cases} y_{ij}^{*}, & \text{if } y_{ij}^{*} > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Multilevel Tobit Model (Cont'd)

- level-2 오차항

$$u_{0j} \sim N(0, \sigma^2_{u0})$$

$$u_{1j} \sim N(0, \sigma^2_{u1})$$

- level-1 오차항

$$\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

- cross-level covariance=0 으로 가정한다.

Multilevel Tobit Model (Cont'd)

- Random intercept u_{0j} : 상위레벨(level-2)에 따라 상수항이 서로 이질적이다.
- Random slope u_{1j}: 상위레벨(level-2)에 따라 x_{ij}가 y_{ij}에 미치는 효과가 서로 이질적이다.

Multilevel Tobit Model (Cont'd)

- Uncensored Probability: $\Pr(y_{ij} > 0 | x_{ij}, u_{0j}, u_{1j})$

  - 입원사건이 발생할 확률로 이해

$$\Pr(y_{ij}^{*} > 0) = 1 - \Phi(\delta)$$

where $\delta \equiv -\frac{\alpha+\beta x_{ij}+u_{0j}+u_{1j}x_{ij}}{\sigma_{\epsilon}}$, $\Phi()$는 표준정규분포의 CDF

Multilevel Tobit Model (Cont'd)

- Truncated Mean : 입원사건이 발생했을 때 평균적 입원의 료비

$$E[y_{ij}|y_{ij}^{*} > 0] = \alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij} + \sigma_{\epsilon} \times \lambda(\delta)$$

where $\lambda(\delta) = \frac{\phi(\delta)}{1-\Phi(\delta)}$

Multilevel Tobit Model (Cont'd)

- Censored Mean: 입원사건 발생여부를 알수 없는 상황에서 평균적 입원의료비

$$E[y_{ij}] = \Pr(y_{ij}^* > 0) \times E[y_{ij}|y_{ij}^* > 0]$$

- Truncated Mean이 Censored Mean보다 항상 큰 값이 된다는 것을 예상할 수 있다.

- 추정계수 β는 입원사건 발생 확률과 평균 입원의료비에 같은 방향으로 영향을 미친다.
  $\Longrightarrow$ 이러한 제약을 완화한 모형이 Two-part Model 또는 Heckman Sample Selection Model

Multilevel Tobit Model (Cont'd)

- 종속변수: 1년동안 입원의료비 로그값

```
. qui metobit lin_mexp i.female IND_age cdcount, ll(0) nolog
. qui metobit lin_mexp i.female IND_age cdcount || pidwon:, ll(0) nolog
. metobit lin_mexp i.female IND_age cdcount || pidwon:cdcount, ll(0) cov(ind) nolog
Mixed-effects tobit regression                  Number of obs     =      5,000
                                                  Uncensored      =        838
Limits: Lower = 0                                 Left-censored   =      4,162
        Upper = +inf                              Right-censored  =          0
Group variable: pidwon                          Number of groups  =        950
                                                Obs per group:
                                                              min =          1
                                                              avg =        5.3
                                                              max =          7
Integration method: mvaghermite                 Integration pts.  =          7
                                                Wald chi2(3)      =      63.47
Log likelihood = -5160.217                      Prob > chi2       =     0.0000
```

lin_mexp	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.female	-1.981599	1.133548	-1.75	0.080	-4.203312	.2401143
IND_age	.3085329	.0854246	3.61	0.000	.1411038	.4759621
cdcount	1.490705	.222645	6.70	0.000	1.054329	1.927081
_cons	-48.63995	6.473619	-7.51	0.000	-61.32801	-35.95189
pidwon						
var(cdcount)	.5299696	.581179			.0617732	4.546755
var(_cons)	76.35331	18.51163			47.47409	122.8002
var(e.lin_mexp)	433.1829	27.64502			382.2514	490.9006

```
LR test vs. tobit model: chi2(2) = 75.40                  Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

Multilevel Tobit Model (Cont'd)

Q9) left-censored 관측치의 비율은 얼마인가?

Q10) cdcount 추정계수=1.490에 대한 해석은?

- 같은 가구원 내에서 시점 간 입원의료비의 상관계수

```
. estat icc
Conditional intraclass correlation
```

Level	ICC	Std. err.	[95% conf.	interval]
pidwon	.1498486	.0315284	.0978933	.2225751

```
Note: ICC is conditional on zero values of random-effects covariates.
```

Q11) zero values of random-effects covariates는 무슨 의미인지?

Multilevel Tobit Model (Cont'd)

- Prediction : $\Pr(y_{ij}^* > 0)$

```
predict re*, reffects

predict yhat1 , pr(0,.)
predict yhat2 , pr(0,.) conditional(fixed)
```

- 개인 이질성을 포함한 입원사건 발생확률과 개인 이질성=0으로 간주한 입원사건 발생확률을 계산한다.

Multilevel Tobit Model (Cont'd)

- Prediction : $E(y_{ij}|y_{ij}^* > 0)$

```
predict yhat3 , e(0,.)
predict yhat4 , e(0,.) conditional(fixed)
```

- 개인 이질성을 포함한 평균적 입원의료비와 개인 이질성=0으로 간주한 평균적 입원의료비를 계산한다.
- 입원사건이 발생되었다는 조건 하에서 입원의료비 $\Longrightarrow$ truncated expectation

Multilevel Tobit Model (Cont'd)

- Prediction : $E(y_{ij})$

```
predict yhat5 , ystar(0,.)
predict yhat6 , ystar(0,.) conditional(fixed)
```

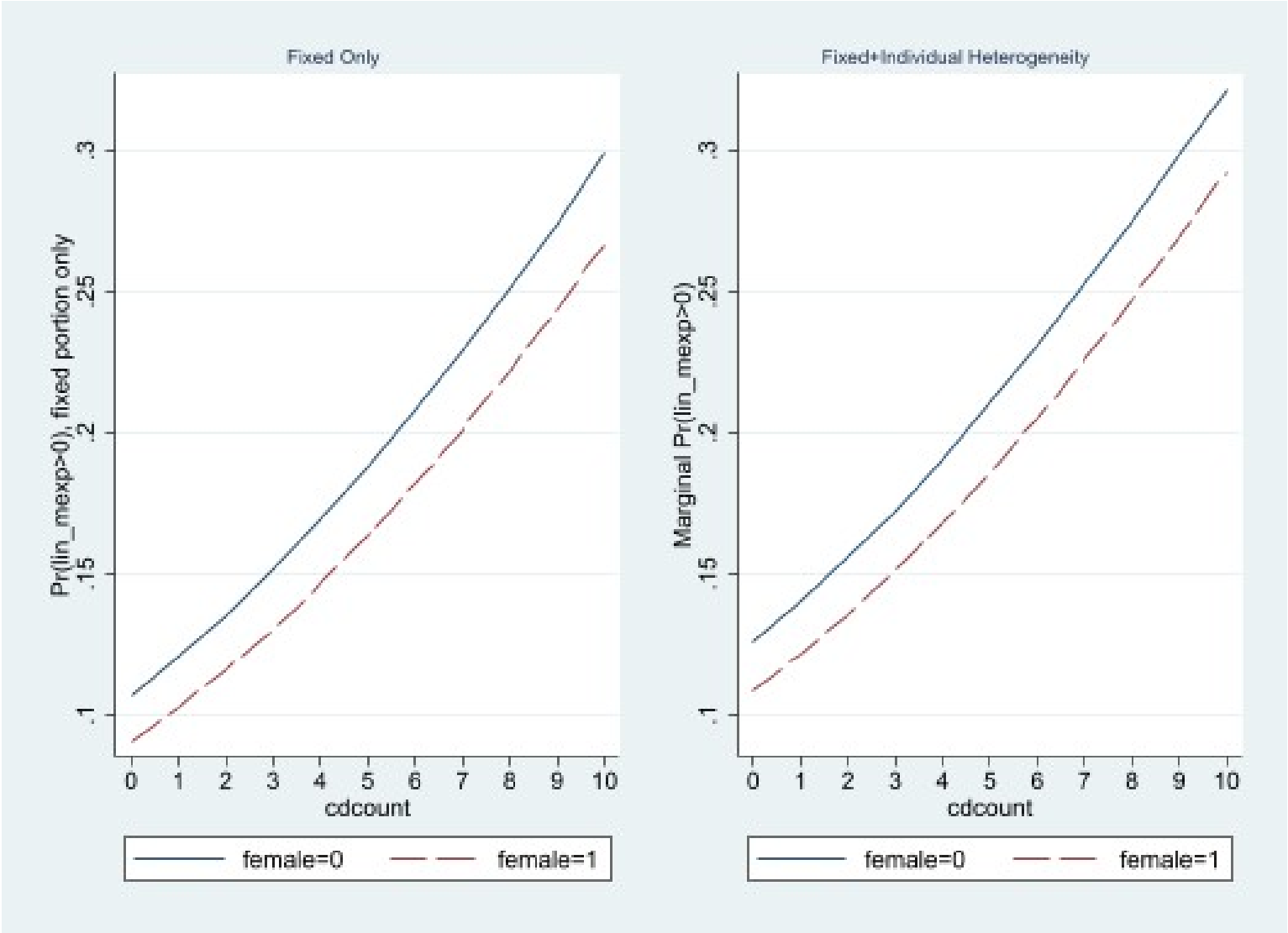
- 입원사건 발생 조건이 주어지지 않은 상태에서 입원의료비
  $\Longrightarrow$ censored expectation

Multilevel Tobit Model (Cont'd)

- Prediction Graph : Fixed Part만 이용한 예측

```
margins female, at(cdcount=(0(1)10)) atmeans ///
predict(pr(0,.) conditional(fixed)) noatlegend

marginsplot, noci recast(line) plot2opts(lpattern(longdash))///
name(graph1, replace) title("Fixed Only", size(small))

margins female, at(cdcount=(0(1)10)) atmeans ///
predict(pr(0,.) marginal) noatlegend

marginsplot, noci recast(line) plot2opts(lpattern(longdash)) ///
name(graph2, replace) title("Fixed+Individual Heterogeneity",size(small))

graph combine graph1 graph2 , ycommon
graph export "fig11.jpg", as(jpg)  replace
```

Multilevel Tobit Model (Cont'd)

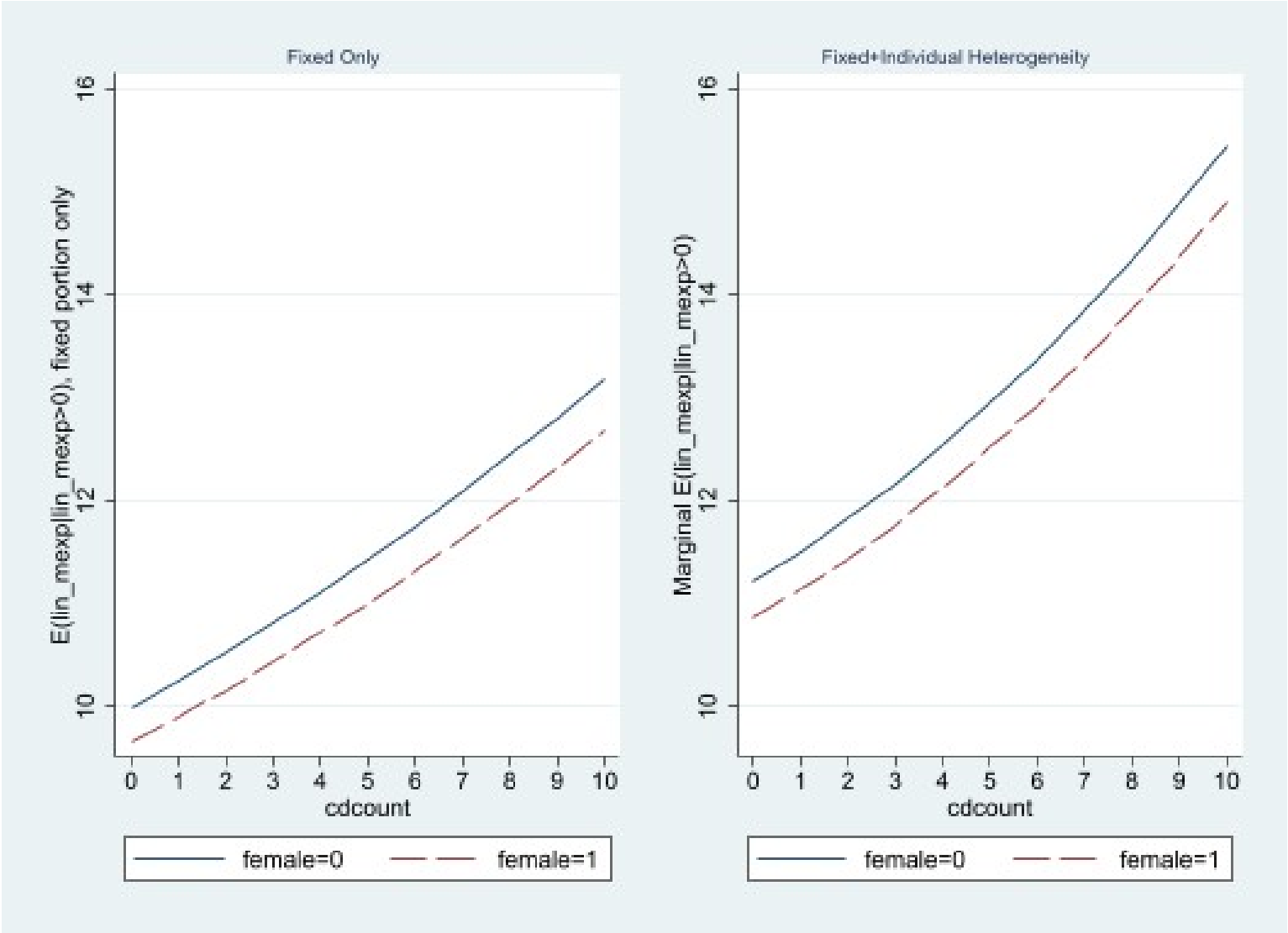


Multilevel Tobit Model (Cont'd)

- Prediction Graph : Fixed Part+individual-specific heterogeneity로 계산한 예측

```
margins female, at(cdcount=(0(1)10)) atmeans predict(e(0,.) ///
conditional(fixed)) noatlegend

marginsplot, noci recast(line) plot2opts(lpattern(longdash)) ///
name(graph1, replace) title("Fixed Only", size(small))

margins female, at(cdcount=(0(1)10)) atmeans ///
predict(e(0,.) marginal) noatlegend

marginsplot, noci recast(line) plot2opts(lpattern(longdash)) ///
name(graph2, replace)  ///
title("Fixed+Individual Heterogeneity",size(small))

graph combine graph1 graph2 , ycommon
graph export "fig12.jpg", as(jpg)  replace
```

Multilevel Tobit Model (Cont'd)



참석해 주셔서 감사드립니다

Thank You !