The Korean Journal of Stata
제5권 1호 2018. 6. pp.31-46

Revisiting Panel Data Analysis (1)
: Stata와 R 코딩

민 인 식*
(경희대학교 경제학과)

< 요 약 >

본 논문에서는 패널데이터 분석의 기초적인 내용을 Stata 15버전과 R 통계패키지를 활용하여 리뷰하고자 한다. 패널분석 초심자를 위해 패널데이터의 개념과 구조에 대한 설명으로 시작한다. 더 나아가 기초통계 분석과정을 설명한다. 예제 데이터는 한국노동패널(Korean Labor and Income Panel Survey: KLIPS) 1차 ~ 19차 데이터를 활용한다. Stata 명령문을 주로 설명하며 R 통계패키지 언어를 이용해서도 유사한 결과를 얻을 수 있다는 것을 보여주고 있다.

주제어: KLIPS, Stata, R, 패널데이터

---

* 교신저자, E-mail: imin@khu.ac.kr
* 본 논문은 2018년 7월 6일(금) 노동패널 자료설명회 워크샵에서 발표될 내용의 일부분이다.

1. 패널데이터 개요와 유형

패널데이터(panel data)는 멀티레벨(multilevel data)의 한 종류라고 말할 수 있다. 멀티레벨 데이터는 상위레벨(level 2)과 하위레벨(level 1)의 two-level 구조로 되어 있는 경우가 대표적인 예이다. 상위레벨 내에 속하는 하위레벨 관측치로 구성되어 있다. 학교(상위레벨)에 속한 학생(하위레벨)들을 예로 들 수 있다. 패널데이터는 상위레벨이 개체(subject)이고 하위레벨이 그 개체를 관찰한 시점(time)으로 구성된 two-level data이다. 그림 1은 상위레벨은 가구(household)이고 하위레벨은 가구레벨을 조사한 시점(wave)으로 구성된 패널데이터 예시이다.

그림 1. 패널데이터 예

그림 2에서는 KLIPS(Korean Labor and Income Panel Survey)에서 구축한 가구레벨 패널데이터의 일부분을 보여준다. 패널그룹을 정의하는 가구id(hhid) 변수와 시간 변수(wave)가 반드시 데이터 내에 존재해야 한다.[1)]

그림 2. 가구레벨 패널데이터

	hhid	wave	year	h_debt_total	h_hage	h_hmarital	h_hsex	h_inc_total
1	1	15	2012	.	71	3	2	3856
2	1	16	2013	.	72	3	2	1180
3	1	17	2014	.	73	3	2	4440
4	1	18	2015	.	74	3	2	4513
5	1	19	2016	.	75	3	2	3760
6	2	15	2012	.	58	3	2	2235
7	2	16	2013	.	59	3	2	920
8	2	17	2014	.	60	3	2	936
9	2	18	2015	.	61	3	2	900
10	2	19	2016	.	62	3	2	1162
11	4	15	2012	8300	42	2	1	5760
12	4	16	2013	2000	43	2	1	6960
13	4	17	2014	.	44	2	1	6960
14	4	18	2015	.	45	2	1	6960
15	4	19	2016	.	46	2	1	6960

1) 패널데이터가 아닌 멀티레벨 데이터인 경우에는 상위레벨을 지정하는 *id* 변수만 있어도 된다.

그룹 id(hhid)는 $i = 1, 2, 3, \cdots, n$로 가구 수는 n개이고 시간 id(wave)는 $t = 1, 2, 3, \cdots, T$로 각 가구가 T년씩 조사되었다고 가정하면 전체 표본의 수는 $N = n \times T$로 계산할 수 있다. 패널 그룹 i내 시간관측치 구성에 따라 패널데이터는 균형패널(balanced panel)과 불균형패널(unbalanced panel)로 구분할 수 있다. 균형패널은 각 패널그룹이 T개씩 관측치를 가지고 있는 경우이다. 그림 2에서 확인할 수 있듯이 hhid=1, 2, 4번은 모두 wave=15~19까지 4년씩 응답한 결과로 구성되어 있다. 불균형패널은 특정 시점에서는 모든 패널그룹 i가 조사된 것은 아니고 제외된 패널그룹이 있다. 따라서 불균형패널에서는 표본의 수는 다음과 같이 쓸 수 있다. 아래의 식에서 T_i는 패널그룹 i가 가지는 시간관측치의 수로 정의한다.

$$N = \sum_{i=1}^{n} T_i$$

unbalanced panel:

person	year	income	age	sex
1	2001	1600	23	1
1	2002	1500	24	1
2	2001	1900	41	2
2	2002	2000	42	2
2	2003	2100	43	2
3	2002	3300	34	1

시간 갭(time gap)을 기준으로 패널데이터 구조를 나누면 시간 갭이 있는 경우와 없는 경우로 구분할 수 있다. 패널그룹 내 시간관측치가 연속적으로 존재하면 시간 갭이 없는 패널데이터이다. 시간 갭이 있는 패널데이터의 예는 특정 가구가 조사응답을 건너뛰었다가 다시 패널에 복귀하는 경우이다. 그림 3의 예제에서 hhid=11은 wave=18에서 응답하지 않고 wave=19 시점에 다시 패널에 복귀하였다는 것을 알 수 있다. 시간 갭이 있는 패널데이터에서 차분 변수를 만들 때 유의해야 한다. 소득의 1차 차분 변수는 다음과 같이 정의할 수 있다. 시간 갭이 없는 hhid=10은 차분된 값이 4번 계산된다. 그러나 시간 갭이 있는 hhid=11은 wave=19에서 차분값이 결측치(missing value)가 된다. hhid=11에서 19차년도 income에서 17차년도 income을 차분하는 실수를 할 수 있음에 유의해야 한다.

$$\Delta\, income_{it} = income_{it} - income_{it-1}$$

그림 3. 시간 갭이 있는 패널데이터

	hhid	wave	h_inc_total	dincome
1	10	15	1000	.
2	10	16	600	-400
3	10	17	1030	430
4	10	18	1220	190
5	10	19	1790	570
6	11	15	2510	.
7	11	16	3750	1240
8	11	17	2220	-1530
9	11	19	3900	.

패널그룹의 수 n과 조사시점의 수 T의 상대적 크기를 비교하여 micro panel과 macro panel로 구분할 수도 있다. micro 패널데이터는 $n > > T$로 패널그룹의 수가 각 패널이 가지는 시간관측치의 수보다 훨씬 많은 경우이다. 주로 가구나 가구원 서베이 데이터에서 관찰되는 패널데이터이다. 노동패널과 같은 데이터는 대표적인 micro 패널데이터이다. 반면 macro 패널데이터는 $n < T$로 패널그룹의 수가 관측 시점의 수보다 적은 경우이다. 장기간 축적된 OECD 국가패널데이터의 그 예가 될 수 있다. macro 패널데이터는 시간 흐름에 따른 개체의 변화를 분석하는데 있어서 많은 정보를 가지고 있다. 반면 micro 패널에서는 특정 시점에서 많은 개체 정보를 가지고 있기 때문에 개체 이질성(unit heterogeneity)을 고려하는 분석에 유용하다.

패널데이터를 다루는 통계패키지에서는 주어진 데이터가 "패널데이터"임을 먼저 알려주어야 한다. 가령 Stata에서는 다음과 같이 tsset 또는 xtset 명령어를 사용한다. 그림 4에서는 Stata 코드의 예시를 보여준다. tsset 명령어 다음에 패널그룹그룹 변수와 시간변수를 순서대로 입력한다. 패널그룹 변수가 문자변수(string variable)일 수 없고 반드시 숫자변수(numeric variable)인 경우에만 tsset 명령문 실행이 가능하다. 실행결과를 살펴보면 unbalance(불균형 패널)이면서 시간 갭이 가진 패널데이터임을 확인할 수 있다. xtset을 사용하더라도 같은 결과를 얻는다. xtset은 패널데이터 뿐 아니라 멀티레벨 데이터에서 사용될 수 있는 장점이 있다. 그림 4 예제와 같이 xtset 다음에 상위레벨 변수(hhid)만 지정하는 것도 가능하다.

그림 4. tsset 과 xtset 명령어: Stata

```
. use klips_final, clear
. tsset hhid wave
      panel variable:  hhid (unbalanced)
       time variable:  wave, 15 to 19, but with gaps
              delta:  1 unit
. xtset hhid
      panel variable:  hhid (unbalanced)
```

그림 5에서는 R 코드 예제를 제시한다. plm library에 있는 명령어를 이용해야 한다. pdata.frame 명령어가 Stata에 tsset과 같다고 생각할 수 있다. index 옵션에서 패널그룹변수와 시간변수를 지정한다. pdim 명령어를 통해 패널데이터의 구조와 표본 크기에 대해 이해할 수 있다.

그림 5. 패널데이터 declare: R 코드 예제

```
library(haven)
library(plm)
klips_final<-read_dta(file="klips_final.dta")
klips_final<-pdata.frame(klips_final, index=c("hhid","wave"))

> pdim(klips_final)
Unbalanced Panel: n = 7633, T = 1-5, N = 34322
```

2. 패널데이터 구축 : smart_klips

KLIPS 데이터는 년도별로 조사된 결과를 가구, 가구원 그리고 부가조사 데이터로 구성되어 있다. 년도별 데이터를 병합하여 하나의 패널데이터로 만들기 위해서는 아래와 같이 두 가지 방법을 활용할 수 있다.

(1) Stata에서 append 명령어의 활용

가구레벨 wave=15 ~ 17차년도 데이터에서 h__1406(가구 입주형태) 변수가 포함된 패널데이터를 만드는 예제를 그림 6과 그림 7에서 제시한다. 그림 6에서는 3개 년도 조사데이터를 아래쪽으로 계속 추가시키면 패널데이터가 완성된다는 것을 보여준다. append를 이용하여 패널데이터로 만들기 위해서는 h151406~h171406 변수 이름을 h1406(입주형태)으로 동일하게 미리 코딩해 두어야 한다.

그림 6. append 명령어

wave=15

hhid	wave	h1406
1	15	1
2	15	2
4	15	1
6	15	4
7	15	.
9	15	2
10	15	1

wave=16

hhid	wave	h1406
1	16	1
2	16	2
4	16	1
6	16	4
7	16	.
9	16	2

wave=17

hhid	wave	h1406
1	17	1
2	17	2
4	17	1
6	17	4
7	17	.
9	17	2

wave15 ~ wave17

hhid	wave	h1406
1	15	1
1	16	1
1	17	1
2	15	2
2	16	2
2	17	2
4	15	1
4	16	1
4	17	1

그림 7. append 코드 예제

```
forvalues j=15/17 {
       use klips`j'h, clear
       keep hhid`j' h`j'1406
       ren hhid`j' hhid
       ren h`j'1406 h1406
       gen wave=`j'
       tempfile file`j'
       save `file`j'', replace
}

use `file15', clear
forvalues i=16/17 {
       append using `file`i'', force
}
drop if hhid==.
order hhid wave
sort hhid wave

save klips_h1406, replace
```

(2) Stata에서 smart_klips 명령어 활용

노동패널 홈페이지에서 제공하는 smart_klips 패키지를 다운로드 받은 후 설치한다[2]. 연구자가 원하는 변수와 wave를 선택하여 패널데이터 셋트로 만들 수 있다. 자세한 설치와 사용방법에 대해서는 민인식(2016) 또는 smart_klips 유저가이드를 참고할 수 있다. 그림 7에서 가구주 나이(h_hage)와 입주형태(h1406) 변수를 wave 15 ~ wave 17에서 가져와 패널데이터로 만들는 예제 코드를 보여준다. 그림 7의 append 예제 코드에 비해 간단하게 2줄 코드만을 이용하여 그림 7과 같은 결과를 얻을 수 있다는 장점이 있다.

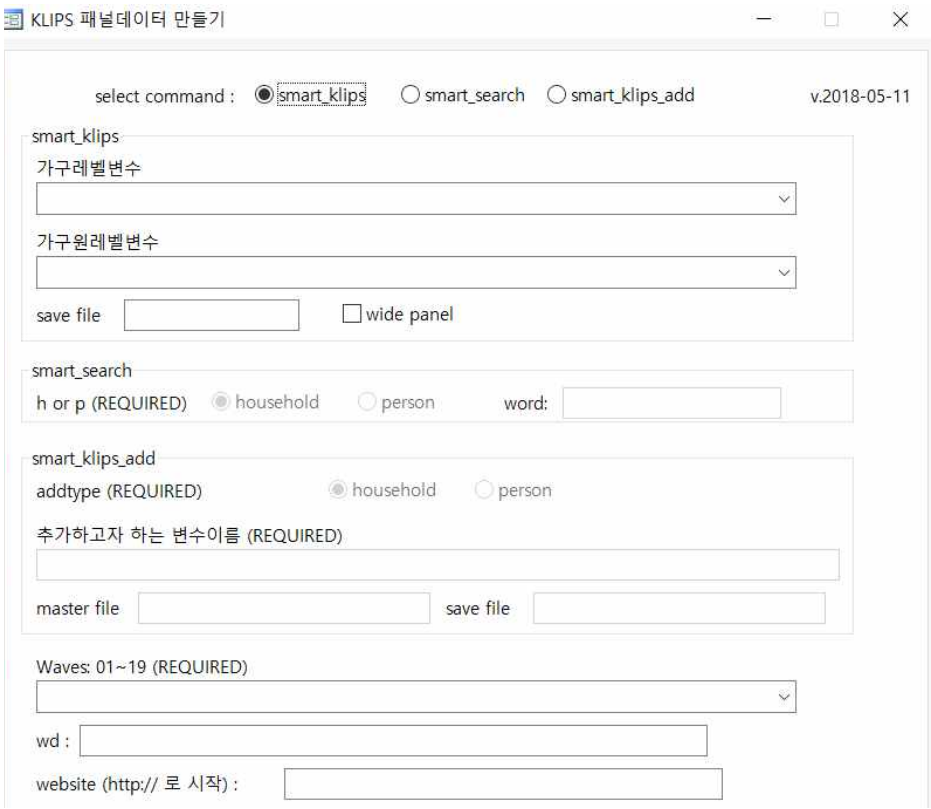
2) 민인식(2018)에서는 smart_klips의 새로운 버전인 smart_klips_v2 패키지를 소개하고 있다.

그림 8. smart_klips 예제 코드

```
smart_klips  h_hage  , wave(15 16 17 )
smart_klips_add  h1406  ,addtype(h)  wave(15 16 17 )
```

Stata 초보자를 위해 smart_klips 패키지는 그림 9와 같은 dialog box를 제공하고 있다. smart_klips, smart_search 그리고 smart_klips_add 명령어를 dialog box를 통해 실행할 수 있다.

그림 9. smart_klips dialog box



R을 이용하여 패널데이터를 구축하는 코드는 따로 소개하지 않는다. 대신 Stata에서 만들어 저장한 패널데이터를 R에서 불러와서 사용하도록 한다.

3. 기초 통계분석[3)]

패널데이터에 있는 x_{it}변수 평균과 분산은 다음과 같이 전체 평균과 각 그룹별 평균(between mean)으로 구분할 수 있다. 식 3의 within mean은 편차(deviation)의 합이기 때문에 항상 0이 나온다.

$$\bar{\bar{x}} = \frac{1}{nT}\sum_{i}^{n}\sum_{t}^{T} x_{it} \quad : \text{전체 평균(overall mean)} \ (\text{식 } 1)$$

$$\frac{1}{n}\sum_{i}^{n}\bar{x}_i \ : \text{between mean} \ (\text{식 } 2)$$

$$\frac{1}{nT}\sum_{i}^{n}\sum_{t}^{T}(x_{it}-\bar{x}_i) \quad : \text{within mean} \ (\text{식 } 3)$$

x_{it}의 분산 역시 평균과 같이 세 가지 유형을 계산할 수 있다. within 분산은 시간불변 변수(time-invariant variable)에서는 $x_{it}=\bar{x}_i$이기 때문에 항상 0이 된다는 것을 확인할 수 있다. between 분산은 패널그룹 간 차이라고 이해할 수 있으면 within 분산은 패널그룹 내에서 관찰된 시점 간 차이로 이해할 수 있다. 따라서 between 분산이 within 분산보다 크다면 패널그룹 간 x변수의 차이가 패널그룹 내 x변수의 차이가 보다 더 크다고 이해한다.

$$\frac{1}{(nT-1)}\sum_{i}\sum_{t}(x_{it}-\bar{\bar{x}})^2 \quad : \text{overall 분산} \ (\text{식 } 4)$$

$$\frac{1}{(n-1)}\sum_{i}(\bar{x}_i-\bar{\bar{x}})^2 \quad : \text{between 분산} \ (\text{식 } 5)$$

$$\frac{1}{(nT-1)}\sum_{i}\sum_{t}(x_{it}-\bar{x}_i)^2 \quad : \text{within 분산} \ (\text{식 } 6)$$

Stata에서 xtsum 명령어를 이용하여 h_inc_total(가구소득) 변수의 평균과 분산 계산 결과를 그림 10에서 보여준다. between mean은 overall mean인 4172.58과 같고 within mean=0이 된다.

3) 좀 더 자세한 패널데이터 기초통계분석 내용은 민인식(2012)을 참고할 수 있다.

그림 10. xtsum 명령어

```
. xtsum h_inc_total
```

Variable		Mean	Std. Dev.	Min	Max	Observations
h_inc_~l	overall	4172.58	3726.747	10	105666	N = 33857
	between		3253.134	10	88200	n = 7554
	within		2055.912	-22247.02	86893.38	T-bar = 4.482

특정 가구에서 관찰한 소득은 시점 간 양의 상관관계가 존재할 수 있다. 이를 intra-class correlation, 즉 패널그룹 내 계열 상관계수(serial correlation within a group)라고 부른다. i 그룹 내 t 시점과 s 시점의 상관계수는 다음과 같이 주어진다. 식 7의 계열상관계수는 t와 s의 차이에 의존하지 않는다고 가정한다. 임의의 두 시점 간 상관계수는 ρ로 동일하다는 equi-correlation으로 가정하고 계산한다.

$$corr(x_{it}, x_{is}) = \frac{T\sigma_b^2 - \sigma_w^2}{T\sigma_b^2 + (T-1)\sigma_w^2} \quad \text{(식 7)}$$

Stata에서는 icc 또는 xtreg 명령어를 이용하여 intra-class correlation을 계산할 수 있다. 그림 11에서는 icc 명령문 실행결과를 보여준다. individual ICC=0.60을 intra-class correlation으로 해석한다. 가구 내 시점 간 가구소득의 상관계수는 60%로 양의 상관관계가 강하게 나타난다.

그림 11. intra-class correlation

```
. icc h_inc_total hhid
 Random effects: hhid                Number of targets =      5863
                                     Number of raters  =         5

------------------------------------------------------------------
         h_inc_total |       ICC       [95% Conf. Interval]
---------------------+--------------------------------------------
          Individual |   .6017276       .5907317    .6126975
             Average |   .8830987          .8783    .8877641
------------------------------------------------------------------
F test that
  ICC=0.00: F(5862.0, 23452.0) = 8.55          Prob > F = 0.000
```

범주형 변수(categorical variable)의 기초 통계량은 빈도(frequency)를 계산하여 보여줄 수 있다. 횡단면 데이터에서 tab(Stata) 또는 table(R) 명령어를 사용할 수 있지만 패널데이터에서는 frequency table을 계산하기 위해 xttab (Stata) 명령어를 사용한다. 그림 12에서는 h1406(주택 점유유형)에 대한 xttab 명령문 실행결과를 보여준다. Overall Freq.는 전체 표본에서 각 범주의 빈도를 계산한다. Between Freq.는 가구 기준으로 계산한 빈도이다. 가령 1(자가)=4946은 전체 7633가구 중에서 4946가구는 조사 기간(wave15~wave19) 동안 1번이라도 자가에 거주한 적이 있다는 의미이다. 전세경험(h1406=2)이 있는 가구 수는 2039가구이다. Within Percent는 가구 i 의 시점 내에서 범주 변화를 측정한 값이다. 자가(h1406=1) 경험이 있는 가구는 전체 조사 기간의 88.9%기간동안 자가에 계속 거주하였다. 조사기간이 5년(wave15~wave19)이면 5년 x 0.889=4.45년이 된다. 4.45년 동안 자가에 계속 거주해 왔다고 해석한다. Within percent 값이 100%에 가까울수록 그 범주를 지속적으로 유지할 가능성이 크다고 이해할 수 있다.

그림 12. xttab 명령어

```
. xttab h1406
```

	Overall		Between		Within
h1406	Freq.	Percent	Freq.	Percent	Percent
1	20575	59.97	4946	64.80	88.94
2	6330	18.45	2039	26.71	71.61
3	5389	15.71	1674	21.93	75.78
4	2017	5.88	760	9.96	66.54
Total	34311	100.00	9419	123.40	81.04
			(n = 7633)		

패널데이터는 시간흐름에 따라 조사되는 구조이기 때문에 전이확률(transition probability)을 계산할 수 있다. 전이확률은 식 8과 같이 정의된다. $t-1$ 시점에서 k 상태에 있는 가구가 다음 기인 t 시점에서는 s 상태로 바뀌는 확률이다. Stata에서는 xttrans 명령어를 이용하여 전이확률을 계산할 수 있다.

$$\Pr(x_{it} = s \mid x_{it-1} = k) \quad \text{(식 8)}$$

그림 13에서는 h1406(입주형태) 변수의 4개 범주에 대한 전이확률 계산결과를 보여준다. $t-1$ 시점에 자가(h1406=1)였던 가구는 다음 기인 t 시점에서 여전히 자가일 확률은 97%=15564/16031이다. 자가에서 전세로 이동한 가구는 1.7%이다. 자가 지속성 비율이 매우 높다는 것을 이해할 수 있다. 또 다른 측면에서 해석하면 전이확률은 범주형 변수의 계열상관(serial correlation)이라고 볼 수 있다.

$t-1$시점에서 전세인 가구는 다음 기인 t시점에서 자가로 이동한 가구 비율은 9.6%=483/5001이다.

그림 13. xttrans 명령어

```
. xttrans h1406, freq
```

입주형태	입주형태 1	입주형태 2	입주형태 3	입주형태 4	Total
1	15,564 97.09	272 1.70	109 0.68	86 0.54	16,031 100.00
2	483 9.66	4,231 84.60	227 4.54	60 1.20	5,001 100.00
3	161 3.93	180 4.39	3,659 89.29	98 2.39	4,098 100.00
4	99 6.41	82 5.31	135 8.74	1,229 79.55	1,545 100.00
Total	16,307 61.13	4,765 17.86	4,130 15.48	1,473 5.52	26,675 100.00

그림 13의 xttrans 결과 해석에서 주의할 점이 있다. 패널데이터가 시간 갭을 가지고 있을 수 있다. 가령 특정 가구가 15, 17, 18, 19(wave=16 제외)와 같이 4개 시점에서 조사된다면 xttrans는 wave=17의 $t-1$시점을 wave=15로 간주하는 문제가 생길 수 있다. 식 8과 같이 t시점과 $t-1$시점과 전이확률을 정확하게 계산하기 위해서는 xttrans 명령문을 실행하기 전에 tsfill 명령문을 먼저 실행하여 시간 갭을 결측치로 채워준 후 xttrans 명령어를 실행한다. 그림 13와 그림 14의 xttrans 실행결과는 큰 차이가 없다는 것을 확인할 수 있다.

그림 14. tsfill과 xttrans 명령문

```
. tsfill
. xttrans h1406
   입주형 |                 입주형태
      태 |         1          2          3          4 |     Total
---------+----------------------------------------------+----------
        1 |     97.24       1.63       0.64       0.49 |    100.00
        2 |      9.52      84.93       4.43       1.11 |    100.00
        3 |      3.75       4.15      89.75       2.35 |    100.00
        4 |      6.13       5.14       8.56      80.17 |    100.00
---------+----------------------------------------------+----------
    Total |     61.28      17.79      15.46       5.47 |    100.00
```

4. Graphical Analysis

패널데이터는 변수의 횡단면 특성과 시계열 특성을 동시에 포함하고 있기 때문에 시간 흐름에 따른 변화를 그래프로 보여줄 수 있다. 본 장에서는 패널데이터에서 사용할 수 있는 그래프 관련 명령어를 소개한다.

패널그룹의 년도별 소득변화를 파악하기 위해 패널 라인그래프를 작성할 수 있다. Stata에서 xtline 명령어를 사용한다. xtline 명령문에서 y축에 표현하고자 하는 변수만 지정하면 된다. x축은 자동으로 시간 변수가 지정된다. 예제 데이터에 가구 id가 매우 많기 때문에 hhid<10인 가구에 대해서만 그래프를 작성한다.

그림 15. xtline 명령어: separate graphs

그림 16에서는 xtline 명령문에서 overlay 옵션을 사용한다. 한 평면에 5개의 line 그래프를 동시에 표현한다. 따라서 시간에 흐름에 따라 가구소득 변화(within-group)뿐 아니라 가구 간(between-group) 소득수준을 비교할 수 있다. 그림 15를 통해 가구소득 변수는 가구 내에서 시간에 따라 변화가 있고 가구 간 소득격차도 존재한다는 것을 알 수 있다.

그림 16. xtline 명령어: overlaid graph

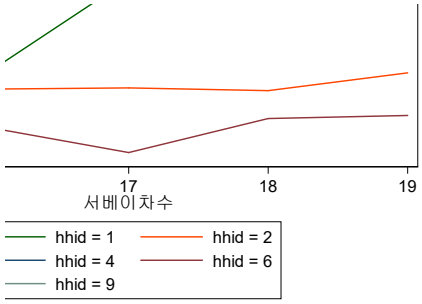
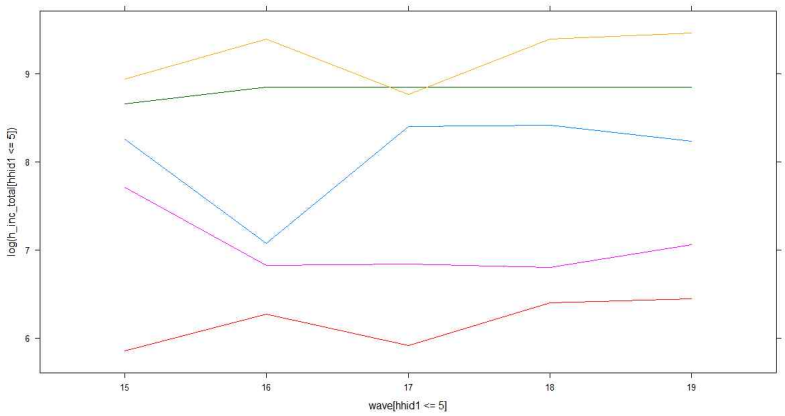


그림 17에서는 그림 16와 같은 그래프를 그리기 위해서 R 코드를 제시한다. lattice library의 xyplot 명령문을 사용한다.

그림 17. 패널 라인그래프 : R 코드

```
library(lattice)
klips_final$hhid1<-as.numeric(klips_final$hhid)

xyplot(log(h_inc_total[hhid1<=5])~wave[hhid1<=5],
group=hhid1, type="l", data=klips_final)
```



로그 가구소득(lincome)과 연령(hage) 변수의 within-variation의 관계를 파악하고자 그림 18을 제시한다. x축과 y축에 해당하는 변수는 식 9-1과 식 9-2와 같이 within-transformation을 적용한 결과이다. Stata에서 xtdata , fe 명령어를 이용하면 원 변수를 within-transformation 변수로 만들어 준다. fitted line은 y변수와 x변수의 단순 선형회귀분석 결과이다. within-transformation 변수 간 선형관계가 분명하다면 가구 내(패널그룹 내)에서 가구주 나이가 증가하면 가구소득이 증가한다고 해석할 수 있다.

$$y = lincome_{it,fe} = (lincome_{it} - \overline{lincome}_i + \overline{\overline{lincome}}) \quad \text{(식 9-1)}$$

$$x = hage_{it,fe} = (hage_{it} - \overline{hage}_i + \overline{\overline{hage}}) \quad \text{(식 9-2)}$$

그림 18. 산포도와 fitted line: within-transformation 변수

```
snapshot save
global snum=r(snapshot)
xtdata, fe clear
twoway (scatter lincome h_hage,msize(vsmall)) (lfit lincome h_hage)
```

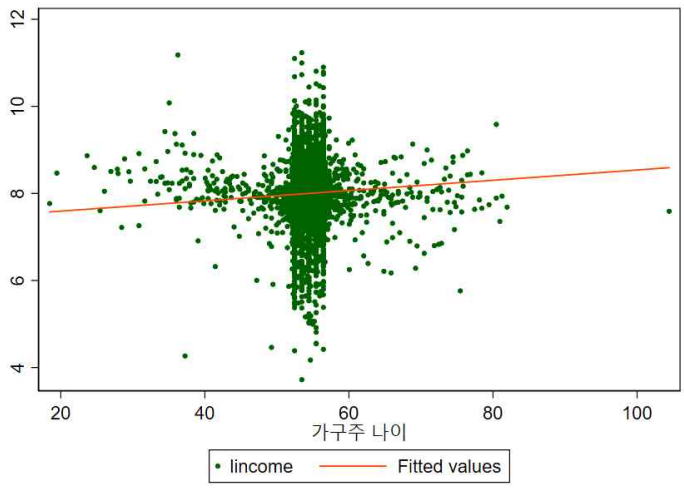


그림 19에서는 between-transformtion 변수를 이용하여 산포도와 fitted line 그래프를 제시한다. y축에 해당하는 변수는 로그 가구소득의 between-transformation 변수이고 x축에 해당하는 변수는 가구주 나이의 between-transformation 변수이다. 식 10-1과 식 10-2에 between-transformation 공식을 제시한다. Stata에서 xtdata , be 명

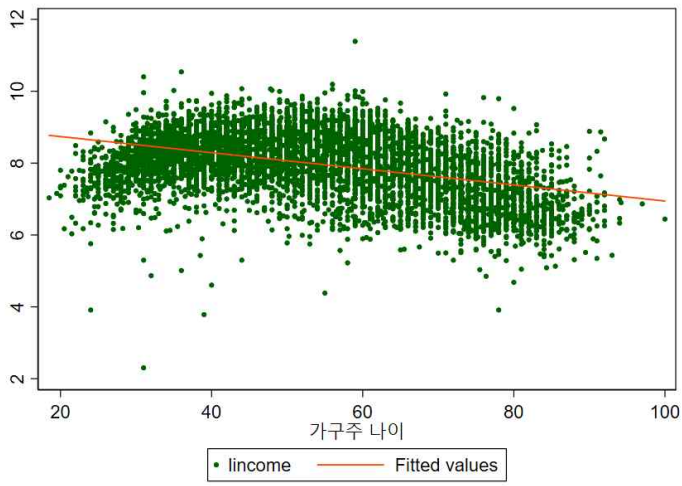
령어를 이용하면 원 변수를 within-transformation 변수로 만들어 준다. 그림 19에서 우하향하는 직선으로 추정된다면 가구주 나이가 높은 가구일수록 가구소득이 낮아진다고 해석할 수 있다. 즉 가구 간(between households) 분석 결과를 제시한 것이다.

$$y = \overline{lincome}_i = \frac{1}{T}\sum_t^T lincome_{it} \quad \text{(식 10-1)}$$

$$x = \overline{hage}_i = \frac{1}{T}\sum_t^T hage_{it} \quad \text{(식 10-2)}$$

그림 19. 산포도와 fitted line: between-transformation 변수

```
snapshot restore $snum
xtdata, be clear
twoway (scatter lincome h_hage,msize(vsmall)) (lfit lincome h_hage)
```



참고문헌

민인식(2016), Stata를 활용한 노동패널 실증분석, 2016년 한국노동패널조사 데이터 설명회 자료집.

민인식(2012), Stata 패널데이터 분석. 제2판. 서울: 지필미디어.

민인식(2018), smart_klips 버전 2.0 소개 및 활용, *The Korean Journal of Stata*, 제5권 1호, pp. 16-23.

The Korean Journal of Stata
제5권 1호 2018. 6. pp.61-73

Revisiting Panel Data Analysis (2)
: Stata와 R 코딩

민 인 식*
(경희대학교 경제학과)

< 요 약 >

본 논문에서는 패널데이터를 이용한 선형회귀모형 추정방법에 대해 설명한다. 모든 패널그룹에서 상수항과 기울기가 동일하다고 가정하는 constant coefficient 모형에 대해 설명한다. 선형회귀모형에 대한 첫 번째 추정량은 오차항에 대해 iid 가정을 하는 POLS 추정에 대해 설명한다. 두 번째 추정량은 오차항의 이분산성, 계열상관, 그리고 contemporaneous correlation을 가정하고 추정하는 패널 GLS 추정에 대해 설명한다. 마지막으로 미시패널데이터 구조에서 주로 사용할 수 있는 패널GEE 추정에 대해 설명한다. 각 추정방법에 대한 Stata와 R 코드를 제시하고 있다.

주제어: KLIPS, Stata, R, 패널 GLS, 패널 GEE

---

* 교신저자, E-mail: imin@khu.ac.kr
* 본 논문은 2018년 7월 6일(금) 노동패널 자료설명회 워크샵에서 발표될 내용의 일부분이다.

1. Linear Regression with panel data

패널데이터는 하나의 개체(subject)를 시간에 따라 반복적으로 조사한 관측치로 구성되어 있다. 따라서 각 관측치는 하첨자 it에 의해 구분된다. i는 패널그룹이고 t는 i그룹 내 시간에 대한 하첨자이다. 주어진 패널데이터에 기초한 선형회귀모형은 다음과 같이 표현할 수 있다.

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \qquad \text{(식 1)}$$
$$\text{where } i = 1, 2, 3, \cdots, n \text{ 이고 } t = 1, 2, 3, \cdots, T$$

식 1 회귀모형에서 주요 가정은 다음과 같다.

가정 1) $\alpha_i = \alpha$이고 $\beta_i = \beta$이다. 즉 모든 패널그룹이 서로 같은 상수항과 기울기 모수 β를 가진다고 가정한다. homogeneity across panel groups about parameters.

가정 2) 불편추정량(unbiased estimator)을 얻기 위 필요한 exogeneity assumption이다. $E(\epsilon_{it} | x_{it}) = 0$ 또는 $cov(x_{it}, \epsilon_{it}) = 0$ 즉 오차항에 대해 외생적인 설명변수를 가정한다.

가정 3) 오차항의 분산 역시 패널그룹 i에 대해 동분산성(homoskedasticity)를 가진다. 또한 패널그룹 i 내에서 serial correlation이 존재하지 않는다. $cov(\epsilon_{it}, \epsilon_{it-1}) = 0$ 으로 가정한다.

식 1의 회귀모형에서 가정 1~ 가정 3이 모두 만족하는 경우에는 모든 i, t를 pooling 하여 OLS 추정한 결과(Pooled OLS)는 일치추정량이면서 효율적 추정량이 된다. 그러나 패널데이터의 속성에 의해 위 가정의 위배가 발생할 수 있고 그에 대한 문제를 해결하기 위해 다양한 panel regression estimation을 활용할 수 있다.

본 논문에서는 KLIPS 데이터를 이용하여 패널데이터를 구축한 후 패널 선형회귀모형을 추정하는 예제를 보여주고자 한다. 본 논문에서 소개하는 추정방법은 어떠한 가정의 위배와 관련이 있고 그러한 가정 위배를 해결하기 위한 접근방식에 대해 설명한다. 각 추정방법에 대한 장단점을 비교하여 연구모형에 필요한 추정량을 선택할 수 있도록 돕고자 한다.

2. Pooled OLS 추정량

식 1에서 제시한 패널 선형회귀모형을 추정하는 첫 번째 방법은 Pooled OLS(POLS) 추정량이다. 가정 1 ~ 가정 3의 모두 만족한다고 가정하면 POLS 추

정량은 효율적이면서 일치추정량이 된다. 모든 패널그룹에 대해 상수항 α와 계수 β가 모두 같다고 가정하기 때문에 constant coefficients 모형이라고 부른다.

먼저 smart_klips 명령어와 12~19차 데이터를 이용하여 실습데이터(klips_example1.dta)를 완성한다. 회귀모형은 임금근로자의 로그임금을 추정하는 모형을 설정한다. y변수는 임금근로자의 월 평균 임금의 로그값이고 설명변수는 가구원의 나이, 나이제곱, 현 직장 근무기간(년), 배우자 유무, 남자여부, 교육수준, 회사규모로 설정한다.

```
  smart_klips h_hsex h_hage h_kidage06 h_hmarital ///
  h_inc_total h_debt_total h_resid_type ///
  p_sex p_age p_edu p_married p_region ///
  p_wage p_firm_size  p_job_begin  p_employ_type , wave(12 13
14 15 16 17 18 19)
  keep if p_employ_type==1  // 임금근로자만 선택한다.

  gen tenure=(ym(year,12)-p_job_begin+1)/12
  replace tenure=. if tenure<0
  recode tenure (60/max=.)

  * y변수 : log(wage)
  gen lwage=log(p_wage)
  recode lwage (0=.)

  * x변수: p_age, p_sex, p_edu, tenure, p_married, p_firm_size
  recode p_edu (1 2 =1) (3 4=2) (5 6=3),gen(edu)
  recode p_firm_size (1 2=1) (3 4=2) (5 6=3),gen(firm_size)
  recode p_married (1 3=0) (2=1),gen(married)
  rename p_age age
  recode p_sex (1=1) (2=0), gen(male)

  save klips_example1, replace
```

위 예제 데이터를 이용하여 POLS 추정량을 얻기 위해 Stata에서 reg 명령어를 사용한다.

그림 1. Stata에서 Pooled OLS 추정결과

```
. use klips_example1, clear
. tsset pid wave
. reg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size

      Source |       SS           df       MS      Number of obs   =    35,362
-------------+----------------------------------   F(9, 35352)     =   4460.48
       Model |  7318.6434         9    813.1826   Prob > F        =    0.0000
    Residual |  6444.96491    35,352  .182308354   R-squared       =    0.5317
-------------+----------------------------------   Adj R-squared   =    0.5316
       Total |  13763.6083    35,361  .389231309   Root MSE        =    .42698

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0695166   .0013388    51.92   0.000      .0668925    .0721407
             |
 c.age#c.age |  -.0008165   .0000143   -56.91   0.000     -.0008446   -.0007884
             |
      tenure |   .0227507   .0003917    58.08   0.000      .0219829    .0235185
      1.male |   .4134277   .0048202    85.77   0.000      .4039799    .4228755
             |
         edu |
          2  |   .1732907   .0079584    21.77   0.000      .1576921    .1888893
          3  |   .4164072    .008657    48.10   0.000      .3994393    .4333751
             |
   1.married |   .0724475   .0056679    12.78   0.000      .0613382    .0835567
             |
   firm_size |
          2  |   .1655044     .00548    30.20   0.000      .1547635    .1762454
          3  |   .3207417   .0063246    50.71   0.000      .3083454    .3331381
             |
       _cons |     3.0388   .0292685   103.83   0.000      2.981433    3.096167
------------------------------------------------------------------------------
```

상수항과 기울기 계수는 모든 가구원에서 서로 같게 주어지는 constant coefficients 모형 추정결과로 해석한다. 따라서 tenure=0.023에 대한 해석은 다음과 같다. 모든 가구원은 개인-시점(년도)과 무관하게 근무 기간이 1년 증가하면 (다른 조건이 일정할 때) 평균적으로 임금이 2.3% 높아진다고 해석한다.

식 1의 오차항 ϵ_{it}가 그룹 내에서 서로 독립이라는 가정을 완화할 수 있다. 즉 $cov(\epsilon_{it}, \epsilon_{it-1}) \neq 0$으로 가정하고 표준오차(standard error)를 다시 계산할 수 있다. 다음과 같이 vce(cluster pid) 옵션을 추가하여 cluster-robust SE를 제시할 수 있다. 추정계수는 그림 1과 같다. 그러나 SE만 새롭게 주어진다. 따라서 t-value와 p-vaue 따라서 유의성에 대한 판단이 달라질 수 있다. 그러나 이러한 cluster-robust SE는 large sample에서만 올바른 SE가 된다. 또한 추정계수는 여전히 오차항의 no serial correlation을 가정하고 계산한 값이기 때문에 효율적 추정량(efficient estimator)이 되지 못하는 단점이 있다.

```
. reg lwage age c.age#c.age tenure i.male i.edu i.married
i.firm_size , vce(cluster pid)

Linear regression                               Number of obs     =     35,362
                                                F(9, 8509)        =    1282.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5317
                                                Root MSE          =     .42698

                                   (Std. Err. adjusted for 8,510 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0695166    .002612    26.61   0.000     .0643964    .0746368
             |
 c.age#c.age |  -.0008165   .0000285   -28.60   0.000    -.0008724   -.0007605
             |
      tenure |   .0227507   .0008202    27.74   0.000     .0211428    .0243586
      1.male |   .4134277   .0094653    43.68   0.000     .3948734    .4319819
             |
      <결과 중략>
------------------------------------------------------------------------------
```

R에서 POLS 추정결과를 얻기 위해서는 그림 2 코드를 사용할 수 있다. OLS 추정량을 얻는 lm 명령어를 사용해도 되고 plm 명령어에서 model="pooling" 옵션을 주어도 같은 결과를 얻는다. cluster-robust SE를 얻기 위해서는 lmtest library를 호출한 후 coeftest 명령문을 작성한다. 지면제약 상 R의 추정결과를 제시하지 않았지만 정확히 Stata 추정결과와 같다는 것을 확인할 수 있다.

그림 2. R에서 POLS 추정결과

```
  library(haven)
  library(plm)
  klips_final<-read_dta(file="klips_example1.dta")
  klips_final<-pdata.frame(klips_final, index=c("pid","wave"))

  klips_final$edu<-as.factor(klips_final$edu)
  klips_final$married<-as.factor(klips_final$married)
  klips_final$male<-as.factor(klips_final$male)
  klips_final$firm_size<-as.factor(klips_final$firm_size)

  # lm 명령어
  pols<-lm(lwage                                                    ~
age+I(age^2)+tenure+edu+married+male+firm_size, data=klips_final)
  summary(pols)

  #plm 명령어
  pols1<-plm(lwage                                                  ~
age+I(age^2)+tenure+edu+married+male+firm_size,
           data=klips_final, model="pooling")
  summary(pols1)

  # cluster-robust SE
  library(lmtest)
  coeftest(pols1,vcov=vcovHC(pols1,type="HC0",cluster="group"))
```

3. Panel GLS 추정량[4)]

앞선 Pooled OLS 추정에서는 패널데이터의 가장 기본적인 특징을 고려하지 않는 문제가 있다. 동일한 개체를 시간에 따라 반복적으로 관찰하는 경우 식 1의 종속변수 y_{it}는 그룹 내 상관관계를 가지고 있을 가능성이 크다. 즉 $cov(y_{it}, y_{it-1}) \neq 0$ for each i 라고 가정하는 것이 일반적이다. Pooled OLS에서는 모든 i, t에 대해서 오차항의 동분산성을 가정하고 있다. 즉 $var(\epsilon_{it}) = \sigma^2_{\epsilon}$ for all i, t 로 가정한다. 그러나 패널그룹 i에 따라 이분산성이 존재할 가능성이 크다. $var(\epsilon_{it}) = \sigma^2_{\epsilon, i}$ for each i로 가정한다.

OLS 추정량 대신 패널데이터의 속성을 고려한 오차항의 이분산성 또는 계열상관(serial correlation)을 고려한 GLS(Generalized Least Squares) 추정량을 사용하면 Pooled OLS보다 더 효율적인 추정량을 얻게 된다.

오차항 가정 1) 모든 i, t에 대해 오차항이 동분산성을 만족하고 no serial correlation을 가정한다. 즉 iid(independent and identically distributed) 오차항을 가정한다. 이 경우는 Pooled OLS 추정량에 문제가 없다.

오차항 가정 2) 오차항이 패널그룹 i에 따라 이분산성을 가진다.
$var(\epsilon_{it}) = \sigma^2_{\epsilon, i}$ for each i

오차항 가정 3) 오차항이 contemporaneous correlation을 가진다. 즉 같은 시점 t에서 서로 다른 그룹 i와 j는 서로 상관관계를 가진다.
$cov(\epsilon_{it}, \epsilon_{jt}) \neq 0$ for $i \neq j$

오차항 가정 4) 오차항이 1계 serial correlation을 가진다. AR(1) 계열상관의 특징은 두 시점이 멀어질수록 계열상관 정도가 점차 줄어든다.
$corr(y_{it}, y_{it-1}) = \rho \neq 0$
AR(1) serial correlation : $\epsilon_{it} = \rho \epsilon_{it-1} + v_{it}$

Stata에서는 오차항 ϵ_{it}에 대해 위 가정 1 - 가정 4를 적용한 패널 GLS 추정량을 얻을 수 있다. xtgls 명령어를 사용한다. 물론 가정 1을 적용한 GLS 추정결과는 Pooled OLS와 같다. Stata에서는 xtgls 명령어를 사용함에 있어 패널그룹의 수에 제약이 있다. 패널그룹의 수가 Stata가 수용할 수 있는 기술적인 행렬 크기(matrix size)를 초과하면 추정결과를 얻을 수 없다. 따라서 아래 예제에서는 개인 id(pid) 값이 30000 이하인 표본으로 한정한다. 이 경우 패널그룹의 수는

4) 본 절에 대한 자세한 설명은 민인식(2015)를 참고하라.

373명이 된다.

panel(iid) 옵션은 오차항 가정 1을 의미한다. 따라서 그림 1의 POLS 추정결과와 같다.[5] panel(hetero) 옵션은 그룹 이분산성을 가정한 GLS 추정결과이다. panel(corr)은 contemporaneous correlation을 가정한 GLS 추정결과이다. 주의할 점은 균형패널(balanced panel)에서만 추정된다. 따라서 본 예제 데이터는 불균형 패널이기 때문에 추정결과를 얻을 수 없다. 마지막 명령문에서 corr(ar1) 옵션은 오차항 가정 4에 해당한다. 추정된 1계 상관계수는 $\hat{\rho} = 0.588$임을 확인할 수 있다. 주의할 점은 corr(ar1) 옵션을 사용하기 위해서는 time gaps이 없는 패널데이터여야 한다. 그러나 본 예제 데이터는 time gaps이 있기 때문에 추정결과를 얻을 수 없다. 다만 force 옵션을 사용하면 time gaps을 무시하고 관측치가 정렬된 순서를 time order로 간주하여 추정결과를 제시해 준다.[6] xtgls는 오차항에 대한 가정이 올바르게 설정한 경우에만 효율적 추정량이 된다는 점에 유의해야 한다. 앞서 언급하였듯이 xtgls는 패널그룹의 수가 많다면 추정에 문제가 있다. 따라서 small n이고 large T인 패널데이터 구조에 적절하다.

```
  use klips_example1, clear
  tsset pid wave
  keep if pid<=30000

  xtgls lwage age c.age#c.age tenure i.male i.edu i.married
i.firm_size , panel(iid)

  xtgls lwage age c.age#c.age tenure i.male i.edu i.married
i.firm_size , panel(hetero)

  xtgls lwage age c.age#c.age tenure i.male i.edu i.married
i.firm_size , panel(corr)

  xtgls lwage age c.age#c.age tenure i.male i.edu i.married
i.firm_size , corr(ar1) force
```

5) 표준오차에서 사소한 차이가 있지만 nmk 옵션을 사용하면 정확히 서로 같다.

6) 가령, 2009년 다음에 2011년 표본이 나오면 2011년 시점에서 $t-1$시점은 2009년으로 간주한다.

R에서 패널 GLS 추정결과를 얻기 위해서는 pggls 명령어를 사용할 수 있다. 그러나 Stata와 같이 오차항 공분산 V 구조를 미리 가정할 수는 없다. 오차항의 공분산 구조를 잔차(residuals)를 이용하여 추정하여 사용한다.

$$\hat{V} = I_n \otimes \hat{\Omega} \text{ where } \hat{\Omega} = \sum_{i=1}^{n} \frac{\hat{\epsilon}_{it}\, \hat{\epsilon}_{it}'}{n} \quad \text{(식 2)}$$

패널 GLS 추정을 위한 R 코드는 아래와 같다.

```
  library(haven)
  library(plm)
  klips_final2<-read_dta(file="klips_example2.dta")
  klips_final2<-pdata.frame(klips_final2, index=c("pid","wave"))
  klips_final2$edu<-as.factor(klips_final2$edu)
  klips_final2$married<-as.factor(klips_final2$married)
  klips_final2$male<-as.factor(klips_final2$male)
  klips_final2$firm_size<-as.factor(klips_final2$firm_size)

  gls1<-pggls(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
data=klips_final2, model="pooling")
  summary(gls1)
```

4. 패널 GEE 추정량

패널 선형회귀모형의 오차항 ϵ_{it}가 iid 오차항이 아닌 경우 3절에서는 패널 GLS 추정량을 사용할 수 있다고 설명하였다. 본 절에서는 오차항의 이분산성, 그룹 내 계열상관과 같이 오차항의 기본 가정이 위배되는 경우 GLS 추정 대신 패널 GEE(Generalized Estimating Equations) 추정량을 활용하는 예를 제시한다. 패널 GEE 추정량은 특히 large n 이면서 small T의 패널데이터 구조에서 주로 활용된다는 점에서 패널 GLS와 차이가 있다. 패널 GEE는 그룹 내 오차항의 계열상관(serial correlation)에 대한 다양한 가정을 할 수 있다는 장점이 있다. 오차항의 이분산성과 contemporaneous correlation 가정에서는 추정할 수 없다는 단점이 있다.

패널 GEE 추정에선 종속변수의 기댓값은 다음과 같이 설명변수의 선형결합의 함수로 표현할 수 있다. $g()$를 link function이라고 부른다. 종속변수 $y_{it} \sim F with\ parameter\ \theta_{it}$와 같이 분포모수 θ_{it}를 가진 특정한 분포 F를 따른다고 가정한다. F를 family distribution이라고 부른다. 패널 선형회귀모형에서는 $g()$는 identity function이 되고 F는 정규분포로 가정하면 된다.

$$g\{E(y_{it})\} = x_{it}\beta \quad (식\ 2)$$
$$또는\ E(y_{it}) = g^{-1}(x_{it}\beta)$$

패널 GEE 추정량에서 오차항이 iid가 아닌 경우 오차항의 구조를 working matrix(상관계수 행렬)로 구조화한다. 특히 패널데이터에서는 그룹 i내의 상관관계 구조를 가정한다. 주로 사용하는 상관관계 구조는 다음 4가지이다.

1) $corr(\epsilon_{it}, \epsilon_{is}) = 0$
그룹 내 계열상관이 없다고 가정한다. iid 가정과 같다.
2) $corr(\epsilon_{it}, \epsilon_{is}) = \rho$ for $all\ \ s \neq t$
현재 시점과 과거 모든 시점과 상관관계는 ρ로 일정하다.
exchangeable correlation이라고 부른다.
3) $corr(\epsilon_{it}, \epsilon_{is}) = \rho^{|t-s|}$ for $all\ \ s \neq t$
현재 시점과 과거 시점의 상관관계는 과거 시점이 멀어질수록 점차 줄어든다.
AR(1) correlation이라고 부른다.
4) $corr(\epsilon_{it}, \epsilon_{is}) = \rho_{ts}$ for $all\ \ s \neq t$
현재 시점과 과거 시점의 상관관계에 대해 미리 특정한 구조를 가정하지 않는다. unstructured correlation이라고 부른다.

Stata에서는 xtreg 또는 xtgee 명령어를 이용하여 패널 GEE 추정결과를 얻을 수 있다. 패널그룹의 수에 대한 제한이 없고 불균형 패널데이터일지라도 추정결과를 얻을 수 있다. 그러나 AR(1) 오차항 구조를 가정하는 경우 time gaps이 있다면 여전히 추정에 문제가 있을 수 있다. xtreg 명령문에서는 pa 옵션을 사용하면 패널 GEE 추정결과를 얻을 수 있다. link function은 identity fucntion이고 family distribution은 gaussian distribution을 사용하고 있다. corr( ) 옵션에서 그룹 내 오차항의 상관관계 구조를 가정한다. 먼저 corr(ind)는 iid 오차항을 가정한다. 따라서 이 결과는 POLS와 같다. corr(exc)는 exchangeable correlation 구조를 가정한다. corr(ar1)은 AR(1) 상관관계를 가정한다. 따라서 time gaps이 있는 패

널그룹은 제외되고 time gaps이 없는 패널그룹만 추정표본(estimation samples)으로 사용한다. 또한 패널그룹 i에서 time period가 1개만 있다면 그 패널그룹 역시 제외된다. 패널 GLS와 같이 force 옵션을 사용하면 time gaps이 있더라도 추정표본에 포함한다. corr(unstructured) 옵션은 오차항의 상관관계 구조를 미리 가정하지 않고 잔차를 이용하여 추정한 구조를 사용한다. xtreg, pa 추정 후 e(R) 행렬에 추정된 상관계수 행렬이 주어진다. 그 결과를 확인하고자 한다면 mat list e(R) 명령문을 이용한다.

그림 3. xtreg, pa 명령문

```
  xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size , pa corr(ind)
  xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size , pa corr(exc)
  xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size , pa corr(ar1)
  xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size , pa corr(ar1)
force
  xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size , pa
corr(unstructured)
   mat list e(R)  // working matrix
```

xtreg, pa 대신 xtgee 명령문을 이용하더라도 그림 3과 같은 결과를 얻을 수 있다. xtgee 명령문은 아래 그림 4에서 제시한다.

그림 4. xtgee 명령문

```
  xtgee lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size
, link(identity) family(gaussian) corr(ind)
  xtgee lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size
, link(identity) family(gaussian) corr(exc)
  xtgee lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size
, link(identity) family(gaussian) corr(ar1)
  xtgee lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size
, link(identity) family(gaussian) corr(ar1) force
  xtgee lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size
, link(identity) family(gaussian) corr(unstructured)
  mat list e(R) // working matrix
```

R에서 패널 GEE 추정결과를 얻기 위해서는 geepack library를 먼저 설치해야 한다. geeglm 명령어를 이용하여 패널GEE 추정결과를 얻을 수 있다. id 옵션에서 패널그룹 변수를 지정하고 corstr 옵션에서 그룹 내 상관계수 구조를 가정한다.

그림 5. R에서 패널GEE 추정

```
library(geepack)
klips_final<-read_dta(file="klips_example1.dta")
klips_final<-pdata.frame(klips_final, index=c("pid","wave"))

klips_final$edu<-as.factor(klips_final$edu)
klips_final$married<-as.factor(klips_final$married)
klips_final$male<-as.factor(klips_final$male)
klips_final$firm_size<-as.factor(klips_final$firm_size)

# 결측치를 미리 제거해준다.
klips_final3 <- na.omit(subset(klips_final,
    select = c(lwage, age, tenure, edu, married, male, firm_size, pid)))
gee1 <-
 geeglm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
             id=pid, data=klips_final3,family=gaussian, corstr="ind")
summary(gee1)
gee2 <-
 geeglm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
             id=pid, data=klips_final3,family=gaussian,
             corstr="exchangeable")
summary(gee2)
gee3 <-
 geeglm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
             id=pid, data=klips_final3,family=gaussian, corstr="ar1")
summary(gee3)
gee4 <-
 geeglm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
             id=pid, data=klips_final3,family=gaussian,
             corstr="unstructured")
summary(gee4)
```

참고문헌

민인식(2015), Stata 패널데이터 분석. 제2판. 서울: 지필미디어.

The Korean Journal of Stata
제5권 2호 2018. 12. pp. 1-16

Revisiting Panel Data Analysis (3)
: Stata와 R 코딩

민 인 식*
(경희대학교 경제학과)

< 요 약 >

본 논문에서는 패널 선형회귀모형으로 varying coefficients 모형을 소개한다. 추정방법으로 고정효과와 확률효과 추정에 대해 자세히 설명한다. 관찰된 설명변수와 group heterogeneity 상관관계 여부에 따라 고정효과 또는 확률효과 추정량이 더 적절할 수 있다. 특히 내생성 문제가 있는 모형에서는 고정효과 추정량만이 일치추정량이 된다. group heterogeneity뿐 아니라 time heterogeneity까지 고려한 two-way effects 모형에 대해서도 설명한다. 각 추정방법에 대한 Stata와 R 코드를 제시하고 있다.

주제어: KLIPS, Stata, R, 고정효과, 확률효과

---

* 교신저자, E-mail: imin@khu.ac.kr
* 본 논문은 2018년 7월 6일(금) 노동패널 자료설명회 워크샵에서 발표될 내용의 일부분이다.

1. Fixed Effects 추정량

패널그룹의 이질성(heterogeneity)을 고려하는 방식으로 오차항 구조를 time-invariant error와 time-varying error로 구성하는 모델링(modelling)을 선택할 수 있다. 식 1에서는 그러한 예를 보여준다.

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it} \quad \text{(식 1)}$$
$$\text{where } i = 1, 2, 3, \cdots, n \text{ 이고 } t = 1, 2, 3, \cdots, T$$

식 1에서 u_i는 종속변수 y_{it}에 영향을 미치는 unobserved factor이고 시간불변 오차항에 해당한다. e_{it} 역시 영향을 미치는 unobserved factor이고 시간가변 오차항에 해당한다. 식 1은 식 2와 같이 바꾸어 쓸 수 있다. 식 2에서는 상수항을 $\alpha + u_i$로 표현할 수 있으며 상수항이 그룹 i에 따라 달라진다는 것에 의해 설명되고 있다. 다만 기울기 모수인 β는 모든 패널그룹에 적용되는 단 하나의 β를 추정하는 것이 목적이다. 식 1(또는 식2)를 error component model 또는 varying coefficients model이라고 부른다. 여기서 varying coefficient는 상수항(intercept term)이 i에 따라 달라질 수 있다는 것을 의미한다.

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

식 1을 추정하는 첫 번째 방법으로 고정효과 추정량(fixed effects estimator)을 사용할 수 있다. "고정효과" 의미는 그룹 이질성(group heterogeneity)에 해당하는 u_i를 확률변수(random variable)가 아니라고 추정해야 할 모수(parameters to be estimated)로 간주하는 의미이다. 고정효과 추정량의 장점은 correlated heteroegeneity를 허용한다. 즉 오차항 u_i와 관찰된 설명변수 x_{it}가 상관관계(즉 내생성: endogeneity)이 있더라도 여전히 α와 β에 대한 일치추정량을 구할 수 있다. 고정효과 추정을 위해 다음 두 가지 접근법을 사용한다. 첫째, u_i를 더미변수로 만들어 직접 그 값을 추정한다. 둘째, u_i를 모형에서 제거하여 직접 추정할 필요가 없는 모형에서 α와 β를 추정하는 것이다.

그룹 이질성을 더미변수로 바꾸어 모형에 포함하는 LSDV(Least Squares Dummy Variables) 추정이 고정효과 추정량이 될 수 있다. 식 1의 u_i를 그룹더미 D_i로 생성한 후 모형에 포함한다.

$$y_{it} = \alpha + \beta x_{it} + \sum_{i=1}^{n-1} \gamma_i D_i + e_{it} \quad \text{(식 3)}$$

식 3에서 time-varying 오차항 $e_{it} \sim iid(0, \sigma_e^2)$로 가정한다. 즉 오차항의 이분산성과 자기상관이 없다고 가정한다. 이러한 가정 하에서 식 3에 대한 OLS 추정량은 불편추정량이면서 효율적인 추정량이 된다. 주의할 점은 패널그룹 n이 너무 많다면 그룹더미를 포함하는 것이 feasible 하지 않을 수 있다.

Stata에서는 그림 1과 같이 *i.* 연산자를 이용하여 그룹더미를 모형에 포함할 수 있다. 예제 데이터에서는 그룹의 수가 9494명으로 매우 많다. 지면제약을 고려하여 그룹의 수를 줄인 후 추정결과를 제시한다. Pooled OLS 추정량을 계산하는 reg 명령어를 사용하고 i.pid를 관찰된 x 변수에 추가하였다. tenure 변수의 추정계수인 0.018은 임금에 영향을 미치는 개인 이질성(그룹 더미변수)을 통제하였을 때 tenure의 1년 증가는 임금을 1.8% 증가시킨다고 해석할 수 있다. 주의할 점은 male 그리고 ed 변수이다. 이 변수들은 time-invariant 변수에 해당한다. 패널고정효과 추정에서는 시간불변 변수의 추정계수는 얻을 수 없다. 그 이유는 시간불변 효과는 모두 u_i에 포함되기 때문에 male와 ed와 개별 변수의 추정계수는 identify 되지 않는다. 다만 그림 1에 제시된 male과 ed 변수의 추정계수를 그대로 report하는 실수를 하지 않아야 한다.[1] LSDV를 추정하는 또 다른 명령문으로는 areg를 사용할 수 있다. areg에서는 absorb(pid) 옵션의 의미는 패널그룹 변수를 더미변수로 포함하게 된다. 앞선 reg 명령문과 같은 결과이지만 male과 ed 변수의 추정계수가 "omitted"로 나타나는 것이 다른 점이다.

1) Stata 추정결과를 살펴보면 male와 ed 변수의 추정계수가 주어진 대신 pid 더미변수 중에서 3개 더미변수가 "omitted"로 나타난다.

그림 1. LSDV 추정: Stata

```
use klips_example2, clear
tsset pid wave
keep if pid<=5000

reg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size i.pid

      Source |       SS           df       MS      Number of obs   =       241
-------------+----------------------------------   F(60, 180)      =     24.79
       Model |  77.8723478         60  1.29787246   Prob > F        =    0.0000
    Residual |  9.42380487        180  .052354471   R-squared       =    0.8920
-------------+----------------------------------   Adj R-squared   =    0.8561
       Total |  87.2961527        240   .36373397   Root MSE        =    .22881
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1098449   .0350957     3.13   0.002     .0405929    .1790968
 c.age#c.age |  -.0009032   .0003589    -2.52   0.013    -.0016115   -.000195
      tenure |     .01847   .0134338     1.37   0.171    -.0080379    .0449779
      1.male |   1.094526   .2092673     5.23   0.000     .6815933    1.507459
         edu |
          2  |   .5721168   .2408094     2.38   0.019     .0969442    1.047289
          3  |   .8666698   .2556837     3.39   0.001      .362147    1.371193
   1.married |  -.1459633   .1142667    -1.28   0.203    -.3714379    .0795112
   firm_size |
          2  |   .1687814   .0875329     1.93   0.055    -.0039412    .3415041
          3  |   .1500092   .1063945     1.41   0.160    -.0599317    .3599501
                              < 이하 생략>

areg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size,
absorb(pid)
```

R 명령문에서는 그림 2와 같이 lm 또는 plm 명령어를 사용할 수 있다. Stata의 i.pid와 같이 factor(pid)를 설명변수에 포함한다. 또한 plm을 사용하는 경우에는 model="pooling" 옵션을 추가한다. R 명령문에서 도출된 추정결과와 Stata의 추정결과는 정확히 일치한다는 것을 확인할 수 있다.

그림 2. LSDV 추정: R

```
library(haven)
library(plm)
klips_final<-read_dta(file="klip_example2_R.dta")
klips_final<-pdata.frame(klips_final, index=c("pid","wave"))
klips_final$edu<-as.factor(klips_final$edu)
klips_final$married<-as.factor(klips_final$married)
klips_final$male<-as.factor(klips_final$male)
klips_final$firm_size<-as.factor(klips_final$firm_size)

lsdv<-lm(lwage ~
 age+I(age^2)+tenure+edu+married+male+firm_size+factor(pid), data=klips_final)
summary(lsdv)

lsdv1<-plm(lwage ~
 age+I(age^2)+tenure+edu+married+male+firm_size+factor(pid),
           data=klips_final, model="pooling")
summary(lsdv1)
```

고정효과 추정방법 중 그룹 이질성을 더미변수화 하는 방법 이외에 u_i를 제거한 모형에서 α와 β를 추정할 할 수 있다. 대표적인 방법은 Within 추정량과 차분(first difference) 추정량이 있다. 본 논문에서는 Within 추정에 대해서 설명한다. u_i는 그룹별 평균을 계산하더라도 여전히 u_i가 되기 때문에 식 3 transformation에서는 사라지게 된다. 식 3에서 u_i는 제외되어 있기 때문에 $cov(x_{it}, u_i) \neq 0$이더라도 식 3에 대한 OLS 추정결과는 α와 β에 대한 일치추정량이 된다. 식 3의 time-varying 오차항 e_{it}에 대해 $e_{it} \sim iid(0, \sigma_e^2)$로 가정하고 OLS 추정결과를 얻게 된다.

$$(y_{it} - \bar{y}_i) = \alpha + \beta(x_{it} - \bar{x}_i) + (e_{it} - \bar{e}_i) \quad \text{(식 3)}$$

그림 3에서는 Within 추정량을 구하는 Stata 명령문을 제시한다. xtreg 명령어에서 fe 옵션을 사용한다. 그림 3의 추정결과에서 확인할 수 있듯이 그림 1의 LSDV 추정결과와 정확히 일치한다. 앞서 설명하였듯이 고정효과 추정량에서는 시간불변 변수인 male과 ed 변수의 추정계수는 따로 identify 되지 않는다. 그림 3

의 맨 아래에 있는 F 검정은 고정효과의 유무에 대한 가설검정 결과이다. 즉 그룹이질성이 존재하지 않는다면 $u_i = 0$ for all i라고 말할 수 있다. 식 3에서 $u_i = 0$인지에 대한 F 검정결과이다. 귀무가설을 기각하면 고정효과 추정량이 적절하고 귀무가설을 기각하지 못하면 그룹 이질성이 없는 Pooled OLS 추정량이 적절하다고 통계적 판단을 한다.

그림 3. Within 추정: Stata

```
xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size, fe
Fixed-effects (within) regression               Number of obs      =       241
Group variable: pid                             Number of groups   =        55
R-sq:                                           Obs per group:
     within  = 0.1745                                         min =         1
     between = 0.0011                                         avg =       4.4
     overall = 0.0003                                         max =         8
                                                F(6,180)           =      6.34
corr(u_i, Xb)  = -0.5919                        Prob > F           =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1098449   .0350957     3.13   0.002     .0405929    .1790968
 c.age#c.age |  -.0009032   .0003589    -2.52   0.013    -.0016115   -.000195
      tenure |     .01847   .0134338     1.37   0.171    -.0080379    .0449779
      1.male |           0  (omitted)
         edu |
          2  |           0  (omitted)
          3  |           0  (omitted)
   1.married |  -.1459633   .1142667    -1.28   0.203    -.3714379    .0795112
   firm_size |
          2  |   .1687814   .0875329     1.93   0.055    -.0039412    .3415041
          3  |   .1500092   .1063945     1.41   0.160    -.0599317    .3599501
       _cons |   1.912402    .828246     2.31   0.022     .2780819    3.546723
-------------+----------------------------------------------------------------
     sigma_u |  .80279296
     sigma_e |    .228811
         rho |  .92486776   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(54, 180) = 14.48                     Prob > F = 0.0000
```

그림 4에서는 Stata의 xtreg, fe에 대응하는 R 명령문을 제시한다. plm 명령어에서 옵션으로 model="within"을 추가한다. 지면제약 상 추정결과를 제시하지 않았지만, Stata와 마찬가지로 male와 ed 변수의 추정계수는 주어지지 않는다. 그림 3의 마지막에 제시된 F검정 결과를 얻기 위해서는 Pooled OLS 추정결과와 Within 추정결과에 대해 F검정을 실시한다. pFtest 명령어를 사용한다. pFtest 명령문 내에 Within 추정결과와 Pooled OLS 추정결과를 저장한 이름을 반드시 순서대로 사용해야 한다.[2] pFtest 검정결과를 그림 3의 F 검정과 사소한 차이가 있다.[3]

그림 4. Within 추정: R

```
within<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
            data=klips_final, model="within")
summary(within)

pols<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
            data=klips_final, model="pooling")
pFtest(within, pols)
```

Pooled OLS 추정량에 비해 고정효과 추정량의 장점은 관찰되지 않는 요인에 의해 내생성(endogeneity)이 존재하더라도 여전히 일치추정량을 얻을 수 있다. 따라서 model mis-specification에 대해 robust한 결과를 얻을 수 있다. 단점은 앞서 보여주였듯이 male과 ed 변수와 같이 시간불변 x변수의 추정계수를 얻을 수 없다.[4] 또한 그룹의 수 만큼 x변수를 추가적으로 사용하는 모형이기 때문에 자유도 감소가 발생한다. 따라서 추정량의 표준오차(standard errors)가 증가하고 유의성이 하락할 가능성이 있다.

2. Random Effects 추정량

식 1에서 제시한 varying-coeffcieints 모형에서 오차항은 다음과 같이 쓸 수 있다. 확률효과 모형에서는 u_i는 고정된 모수가 아니고 확률변수(random variable)로 가정한다. 따라서 패널 그룹별 상수항 $(\alpha + u_i) \sim (\alpha, \sigma^2{}_u)$ 와 같이 확률변수가 된다.[5] 상수항이 확률변수이기 때문에 식 4 모형을 random intercept 모형이라고

2) 순서를 바꾸면 전혀 다른 검정결과가 제시된다.
3) 시간불변 변수인 male와 ed 변수의 계수를 제약조건으로 포함하느냐의 차이이다.
4) 시간불변 변수는 아닐지라도 거의 시간불변 변수에 가까운 설명변수의 표준오차가 매우 커지고 따라서 유의하지 않게 나올 가능성이 있다.

도 부른다.

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it} \quad \text{(식 4)}$$

where $u_i \sim (0, \sigma^2{}_u)$ 그리고 $e_{it} \sim iid(0, \sigma_e^2)$이다. 확률효과 추정량이 일치추정량이 되기 위해서는 uncorrelated heterogeneity, 즉 $cov(x_{it}, u_i) = 0$가 성립해야 한다. 시간불변 오차항과 관찰된 설명변수 간 외생성이 성립해야 한다.

식 4에서 오차항 구조 하에서는 $var(y_{it}) = var(u_i + e_{it}) = \sigma_u^2 + \sigma_e^2$ 로 쓸 수 있다. 오차항 동분산성은 성립하지만 그룹 내 상관관계(serial correlation within a group 또는 intra class correlation)는 0이 아니라는 것을 쉽게 증명할 수 있다. 식 5와 같이 오차항의 serial correlation이 존재하기 때문에 OLS 추정량을 사용할 수 없다.

$$ICC = corr(y_{it}, y_{is}) = \frac{cov(y_{it}, y_{is})}{\sqrt{var(y_{it})}\sqrt{var(y_{is})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} > 0 \quad \text{for all } s \neq t \quad \text{(식 5)}$$

확률효과 추정량의 첫 번째 방법으로 오차항의 상관관계 구조를 가정한 FGLS(Feasible Generalized Least Squares) 추정량을 사용할 수 있다. Swamy and Arora(1972)가 제안한 FGLS 추정은 θ-transformation 모형에서 OLS 추정을 적용한다.

$$(y_{it} - \theta \bar{y}_i) = \alpha(1-\theta) + \beta(x_{it} - \theta \bar{x}_i) + u_i(1-\theta) + (e_{it} - \theta \bar{e}_i) \quad \text{(식 6)}$$

$$\text{where } \theta = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}}.$$

식 6의 오차항 $u_i(1-\theta) + (e_{it} - \theta \bar{e}_i)$은 serial correlation=0이라는 것을 증명할 수 있다 (Swamy and Arora, 1972). $\theta = 0$이면 FGLS 추정량은 Pooled OLS 추정량과 같아지고 $\theta = 1$이면 Within(고정효과) 추정량과 같아지는 것을 쉽게 이해할 수 있다. $\sigma_e = 0$이 되거나 $T \to \infty$이면 $\theta = 1$에 가까워진다. 확률효과 추정에서는 $\theta < 1$이면 시간불변 설명변수의 추정계수를 얻을 수 있는 장점이 있다. FGLS 추정을 사용하는 대신 확률효과 추정량을 얻는 방법은 ML(Maximum

5) 따라서 $\alpha = E(\alpha + u_i)$이기 때문에 α는 group heterogeneity에 대한 모평균(population mean)으로 해석한다.

Likelihood) 추정을 사용할 수 있다. ML 추정을 위해서는 오차항 u_i와 e_{it}에 대한 정규분포(normal distribution)를 가정해야 한다.

확률효과 추정량의 장점은 시간불변 설명변수에 대한 추정계수를 얻을 수 있다. 또한, 고정효과와 달리 자유도 손실이 발생하지 않는다. 따라서 고정효과에 비해 더 많은 표본 수와 변수 정보를 사용하기 때문에 더 효율적인 추정량이 된다. 다만 확률효과 추정량이 일치추정량이 되기 위해서는 uncorrelated heterogeneity를 가정해야 한다.

Stata에서 RE 추정결과를 구하기 위해서는 xtreg 명령문에서 re 옵션을 사용한다. 식 6에서 제시한 Swamy's FGLS 추정결과를 제시해 준다. theta 옵션에서는 θ- transformation을 위한 $\hat{\theta}$을 보여준다. $0 < \hat{\theta} < 1$임을 알 수 있다. male와 ed 변수와 같이 시간불변 설명변수의 추정계수를 얻을 수 있다는 것을 확인할 수 있다. 맨 아래 제시된 rho 값은 ICC의 추정치이다. 그림 3의 고정효과 추정결과에 비해 대부분 추정계수가 통계적으로 유의하다는 차이를 발견할 수 있다. ML 추정결과를 얻기 위해서는 re 옵션 대신 mle 옵션을 사용한다. FGLS와 ML 추정결과는 표본 크기(sample size)가 커질수록 서로 같은 값으로 수렴한다. ML 추정결과의 맨 아래쪽에서는 LR 검정결과가 제시되어 있다. LR 검정의 귀무가설은 다음과 같다.

$$H_0 : var(u_i) = \sigma_u^2 = 0$$

위 귀무가설이 성립하면 식 5의 그룹 내 상관계수(ICC)는 0이 됨을 알 수 있다. 따라서 오차항의 serial correlation이 존재하지 않기 때문에 FGLS 추정 대신 Pooled OLS 추정량을 사용하면 충분하다. LR 검정을 통해 확률효과 추정량을 선택할 것인지 Pooled OLS 추정량을 선택할 것인지에 대한 통계적 결론을 내릴 수 있다.

그림 5. 확률효과 추정: Stata

```
xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size, re theta

Random-effects GLS regression                   Number of obs     =       241
Group variable: pid                             Number of groups  =        55
R-sq:                                           Obs per group:
     within  = 0.1259                                         min =         1
     between = 0.5526                                         avg =       4.4
     overall = 0.4624                                         max =         8
                                                Wald chi2(9)      =     92.04
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =    0.0000
------------------ theta -------------------
  min      5%       median        95%      max
0.4785   0.4785     0.7078     0.7888   0.7888
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0887105   .0210789     4.21   0.000     .0473967    .1300243
 c.age#c.age |  -.0009668    .000217    -4.46   0.000    -.0013921   -.0005416
      tenure |   .0216433   .0068451     3.16   0.002     .0082272    .0350594
      1.male |   .4221465   .1168675     3.61   0.000     .1930904    .6512026
         edu |
          2  |    .269986   .1504729     1.79   0.073    -.0249355    .5649075
          3  |   .5968467    .165715     3.60   0.000     .2720512    .9216422
   1.married |    .009658   .0881612     0.11   0.913    -.1631348    .1824508
   firm_size |
          2  |   .1854374   .0810426     2.29   0.022     .0265968     .344278
          3  |   .1538149   .1003943     1.53   0.125    -.0429543     .350584
       _cons |   2.427399   .4787198     5.07   0.000     1.489125    3.365672
-------------+----------------------------------------------------------------
     sigma_u |  .37440852
     sigma_e |    .228811
         rho |  .72807987   (fraction of variance due to u_i)
------------------------------------------------------------------------------

xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size, mle
nolog
                                <결과 중략>
LR test of sigma_u=0: chibar2(01) = 137.27                Prob >= chibar2 = 0.000
```

확률효과 추정량은 PA 추정량 중에서 exchangeable correlation 가정 결과와 거의 일치한다. 그 이유는 식 5에서 제시된 그룹 내 오차항 상관관계 구조가 정확히 exchangeable correlation 구조와 같기 때문이다. 이론적으로 균형패널(balanced panel)이면 xtreg, pa corr(exc) 결과와 xtreg, re (또는 xtreg, mle) 결과는 서로 일치하고 불균형 패널(unbalanced panel)일지라도 표본 크기가 커지면 PA 추정량과 RE 추정량 결과는 거의 같아진다.

그림 6에서는 확률효과 추정결과를 얻기 위한 R 명령문을 제시한다. plm 명령문에서 model="random" 옵션을 사용하면 Swamy's FGLS(Feasible GLS) 추정결과를 제시한다. 그림 5에서 제시한 $var(u_i) = 0$ 귀무가설에 대한 가설검정은 LR 검정 대신 R에서는 LM(Lagrange Multiplier) 검정결과를 얻는다. plmtest 명령어를 이용하여 Pooled OLS 추정결과와 비교하여 귀무가설을 검정한다. 지면제약상 검정결과를 제시하지 않았지만 p-value는 0.05보다 훨씬 작은 값이다. 따라서 Stata에서 LR 검정과 마찬가지로 귀무가설을 기각한다. Random effects 추정량이 Pooled OLS 추정량보다 적절하다고 판단할 수 있다. 그림 6의 마지막 명령어는 확률효과 추정에 대한 MLE를 얻는 명령문이다. plm library 대신 nlme library를 설치한 후 lme 명령문을 사용해야 한다. random 옵션을 지정하는 것이 중요하다. 또한 결측치를 미리 제거한 데이터 세트를 만든 후에 추정에 사용한다.

그림 6. 확률효과 추정: R

```
re<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
          data=klips_final, model="random")
summary(re)

pols<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
         data=klips_final, model="pooling")
plmtest(pols, effects="individual")

# ML 추정
library(nlme)
klips_final3 <- na.omit(subset(klips_final,
        select = c(lwage, age, tenure, edu, married, male, firm_size, pid)))

fm1    <-    lme(lwage    ~    age+I(age^2)+tenure+edu+married+male+firm_size,
random=~1|pid,  data=klips_final3)

summary(fm1)
```

3. Two-way effects 모형

1절과 2절에서는 그룹이질성(group heterogeneity)만을 모형에 포함하는 one-way FE/RE 모형에 해당한다. 본 절에서는 그룹이질성 뿐 아니라 시간이질성(time heterogeneity) 역시 모형에서 고려하는 것이다.

$$y_{it} = \alpha + \beta x_{it} + u_i + \mu_t + e_{it} \quad \text{(식 7)}$$

식 7에서 μ_t는 t 시점의 모든 패널그룹 i가 서로 같은 값을 갖는다. 즉 관찰되지 않는 시간특성(unobserved time factors)에 해당한다. one-way 모형과 같이 u_i와 μ_t에 대해 고정효과로 간주하거나 확률효과로 간주하고 추정한다. correlated heterogeneity를 허용하는 경우, 즉 $cov(x_{it}, u_i) \neq 0$ 그리고 $cov(x_{it}, \mu_i) \neq 0$인 경우에는 고정효과 추정량이 일치추정량이 된다. 그러나 uncorrelated heteogeneity를 가정하는 경우는 확률효과 추정량이 일치추정량이면서 더 효율적인 추정량이 된다.

이질성 요인을 모두 고정효과로 간주하는 경우에는 1절에서 이미 설명하였듯이 고정효과에 대해 더미변수로 변환한 후 OLS 추정을 하면 된다. $time_t$는 t 시점에 대한 더미변수이다. 식 8에 대한 OLS 추정량이 two-way LSDV 추정량에 해당한다. LSDV 추정대신 within 변환 후 고정효과 추정결과를 얻는 것도 가능하다. 자세한 내용은 민인식·최필선(2015)의 2장을 참고할 수 있다.

$$y_{it} = \alpha + \beta x_{it} + \sum_{i=1}^{n-1} \gamma_i D_i + \sum_{t=1}^{T-1} \delta_t \, time_t + e_{it} \quad \text{(식 8)}$$

식 7의 u_i와 μ_t에 대해 확률변수(random variable)로 가정하면 two-way RE 모형이 된다. 즉 $u_i \sim (0, \sigma_u^2)$ 분포이고 $\mu_t \sim (0, \sigma_\mu^2)$ 분포를 따른다. 오차항의 전체 분산 그리고 serial correlation within a group, contemporaneous correlation은 다음과 같이 계산된다.

$$var(u_i + \mu_t + e_{it}) = \sigma_u^2 + \sigma_\mu^2 + \sigma_e^2$$

$$corr(y_{it}, y_{is}) = \frac{cov(y_{it}, y_{is})}{\sqrt{var(y_{it})}\sqrt{var(y_{is})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\mu^2 + \sigma_e^2} \quad \text{(식 9)}$$

$$corr(y_{it}, y_{jt}) = \frac{cov(y_{it}, y_{jt})}{\sqrt{var(y_{it})}\sqrt{var(y_{jt})}} = \frac{\sigma_\mu^2}{\sigma_u^2 + \sigma_\mu^2 + \sigma_e^2} \quad \text{(식 10)}$$

위 식 9에서 보였듯이 패널그룹 i 내에서 serial correlation 그리고 t 시점에서 comtemporanesou correlation이 0이 아니기 때문에 GLS 추정량 또는 ML 추정량을 사용한다. ML 추정량의 경우에는 u_i, μ_t 그리고 e_{it}에 대해 모두 정규분포를 가정한다. GLS 추정을 위한 θ-transformation의 구체적인 과정에 대해서는 민인식·최필선(2015) 2장을 참고할 수 있다.

Stata에서는 two-way FE 추정량을 구하는 명령어가 따로 없기 때문에 연구자가 직접 패널그룹과 시간에 대해 더미변수를 생성한 후 LSDV 추정량을 얻어야 한다. 또는 xtreg, fe 명령문에서 시간에 대한 더미변수만 추가하는 것도 같은 결과를 가져온다. 그림 7의 i.year 부분이 time hetergoeneity에 해당한다.

그림 7. two-way FE : Stata

```
reg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size i.pid i.year

xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size i.year , fe
```

R에서는 여전히 plm library의 plm 명령문을 여전히 활용할 수 있다. 고정효과 모형이기 때문에 model="within" 옵션을 사용하며 effect="twoways"으로 바꾸어 준다. R에서 group effect만 고려할 때는 effect="individual"이라고 쓰는데 effect 옵션을 사용하지 않으면 자동으로 effect="individual" 옵션이 사용된다. 그림 8에서 두 번째 명령문은 Stata와 유사하게 패널그룹과 시간에 대한 더미변수를 포함해서 two-way FE 모형을 추정하게 된다.

그림 8. two-way FE : R

```
# within transformation
twoway_fe<-plm(lwage ~
 age+I(age^2)+tenure+edu+married+male+firm_size,
            data=klips_final, effect="twoways", model="within")
summary(twoway_fe)

# LSDV 추정 : two-way effects
twoway_fe1<-plm(lwage ~
age+I(age^2)+tenure+edu+married+male+firm_size+factor(pid)+factor(wave),
               data=klips_final, model="pooling")
summary(twoway_fe1)
```

two-way RE 모형에 대한 추정을 위한 Stata 명령문은 xtreg 대신 mixed 명령문을 사용해야 한다. one-way RE에서 설명하였듯이 θ-transformation (GLS 추정량)을 사용할 수도 있고 오차항 분포에 대해 정규분포를 가정하고 MLE 추정결과를 사용할 수도 있다. Stata에서는 MLE 추정결과만 얻을 수 있다.

그림 9. two-way RE : Stata

```
mixed lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size || _all:
R.pid || _all: R.wave, nolog mle
```

그림 10에서는 two-way RE에 대한 R 코드를 제시한다. 주의해야 할 점은 random.method에서 기본 옵션인 "swar"을 사용하는 대신 "walhus"를 지정해야 한다(불균형 패널인 경우). Wallace and Hussain (1969)에서 제시한 θ-transformation을 이용하여 GLS 추정결과를 얻는다.

그림 10. two-way RE : R

```
 twoway_re<-plm(lwage ~
 age+I(age^2)+tenure+edu+married+male+firm_size,
              data=klips_final, effect="twoways", model="random",
random.method = "walhus")
summary(twoway_re)
```

4. Hausman 검정

본 소절에서는 다시 one-way FE 또는 RE 모형으로 돌아간다. 하우스만 검정 (Hausman test)을 이용하여 FE와 RE 모형선택을 위한 가설검정을 실시할 수 있다. 하우스만 검정의 귀무가설은 $cov(x_{it}, u_i) = 0$ 즉 uncorrelated hetereogeneity 이다. 귀무가설을 받아들인다면 RE와 FE 중에서 RE 모형을 선택한다. 그 이유는 uncorrelated heterogeneity에서는 FE와 RE 모두 일치추정량이지만 RE 추정량이 더 효율적인 추정량이기 때문에 RE 추정량을 선택한다. 귀무가설을 기각(reject)한다면 correlated heterogeneity가 존재하고 따라서 FE 추정량만 일치추정량이 되기 때문에 FE 추정량을 선택한다.

하우스만 검정의 기본적인 아이디어는 다음과 같다. 귀무가설이 맞다면 (uncorrelated heterogeneity) FE와 RE 추정치 모두 일치추정량이기 때문에 서로 크게 다르지 않을 것이다. 귀무가설이 기각된다면(즉 correlated heterogeneity라면)

RE 추정치는 FE 추정치와 유의하게 다르다는 것을 확인한다. 검정통계량은 FE와 RE 추정에서 얻은 추정계수의 차이를 공분산 행렬의 차이로 가중하여 계산한다. FE에서는 시간불변 변수의 추정계수를 얻을 수 없기 때문에 시간가변 변수에 대한 추정계수만을 이용하여 하우스만 검정통계량을 계산한다. 하우스만 검정통계량은 근사적으로(asymptotically) $\chi^2(df)$ 분포를 따른다. 여기서 자유도(df)는 시간가변 변수의 수에 해당한다. 또한 하우스만 검정은 FE와 RE 모형(식 7)에서 시간가변 오차항 e_{it}가 iid가정을 만족해야 한다. 만약 iid가정을 만족하지 못한다면 Wooldridge(2002)에서 제시한 robust Hausman test를 활용한다.

Stata에서 하우스만 검정결과를 얻기 위한 코드는 그림 11에서 제시한다. FE와 RE 추정결과를 얻은 후 estimates store 명령어를 이용하여 그 결과를 저장한다. hausman 명령에서 저장한 이름을 FE와 RE 순서대로 입력해야 한다. 검정결과 p-value가 0.05보다 작으면 귀무가설(uncorrelated heterogeneity)을 기각한다. 기각한다면 FE 모형이 RE 모형에 비해 적절하다고 통계적 판단을 한다.

그림 11. 하우스만 검정: Stata

```
xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size, fe
estimates store FE

xtreg lwage age c.age#c.age tenure i.male i.edu i.married i.firm_size, re
estimates store RE

hausman FE RE
* hausman RE FE // 순서를 바꿔서는 안 된다.
```

그림 12에서는 R에서 하우스만 검정을 위한 명령문과 그 결과를 해석한다. 명령문 작성에서 주의할 점은 Stata에서는 저장한 결과를 나열할 때 반드시 FE와 RE 순서로 위치해야 한다. 그러나 R에서는 순서를 다르게 쓰더라도 같은 검정결과를 보여준다.

그림 12. 하우스만 검정: R

```
fixed<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
           data=klips_final, model="within")

random<-plm(lwage ~ age+I(age^2)+tenure+edu+married+male+firm_size,
        data=klips_final, model="random")

phtest(fixed, random)
phtest(random, fixed) # 순서를 바꿔도 같은 결과를 얻는다.
```

참고문헌

민인식(2015), Stata 고급 패널데이터 분석. 서울: 지필미디어.

Swamy, P.A.V.B. and Arora, S.S. (1972) The exact finite sample properties of the estimators of coefficients in the error components regression models, *Econometrica*, 40(2), pp. 261 - 275.

Wallace, T.D. and Hussain, A. (1969) The use of error components models in combining cross section with time series data, *Econometrica*, 37(1), pp. 55 - 72.

Wooldridge, J. (2002) Econometric analysis of cross section and panel data, Cambridge, MA: Massachusetts Institute of Technology.