한국복지패널을 이용한 패널분석모형 산책

한국복지패널 데이터 설명회

2024.04.18

김현식

sochyunsik@khu.ac.kr

경희대학교 사회학과

목차

1: 자료 소개

- 연구 가설
- 종단자료 만들기

연구 가설 1

- 장애인과 비장애인의 노동조합 임금프리미엄 효과 분석
  - 김현식, 2021, 장애인의 노동조합 참여와 임금 차이, 장애와 고용, 31(4), 107-126
  - 장애인 임금 노동자를 무조합원, 비조합원, 조합원으로 구분한 후 임금을 비교하면, 비조합원의 임금은 무조합원의 임금과 별 차이를 보이지 않았으나 조합원은 무조합원에 비해 높은 임금을 받았다

- 연구 자료
  - https://www.koweps.re.kr:442/main.do
  - 한국복지패널 15차를 제외한 8차~18차년도 자료
  - 종속 변수로 사용되는 한달 평균 임금은 8차~18차년도에 측정
  - 혼동 변수 중 일을 시작한 년도에 대한 질문이 15차년도에 없었음
  - 한국복지패널 홈페이지에서 회원 가입 후 간단한 데이터 사용 설문을 진행하고 자료를 다운 받을 수 있음
  - 이 분석에서는 1_18차 결합데이터 파일 중 “koweps_hp01_18_long_240329.sav”를 사용
  - 한국의 많은 자료들이 SPSS 파일로 만들어져 있기 때문에 여기서는 SPSS 파일을 받았다고 가정하고 논의를 진행한다
  - 패널의 유저가이드, 통합설문지, 통합코드북은 웹사이트에서 모두 다운받을 수 있음

연구 가설 2

- 설명 변수: 무조합원, 비조합원(2+3), 조합원

문 6) **(문1)의 ①번 응답자만)** 직장에서 노동조합에 가입되어 있습니까?

① 노동조합이 없음
② 노동조합이 있으나 가입대상이 안됨
③ 노동조합이 있고 가입대상이나 가입하지 않았음
④ 노동조합에 가입하였음

- 종속 변수:
  - 로그 시간당 임금=ln(월 평균 임금/(주당 평균 시간*4.345))

규칙적으로 일한 경우 : 주당 평균 [ ][ ][ ] 시간

일한달의 월 평균 임금 : 월 평균 [ ][ ][ ][ ] 만원

연구 가설 3

- 혼동 변수(confounding variables)
  - 원인 변수와 결과 변수 모두에 영향을 주는 변수로 이 변수들을 통제하지 않으면 원인 변수가 결과 변수에 주는 영향이 왜곡(biased)되어 나타남
  - 장애인 여부
  - 응답년도, 연령, 성별, 교육수준, 혼인상태
  - 근속년수, 5개 권역별 지역구분
  - 사업체 규모나 업종, 직종 변수를 발견하지 못함

종단 자료 만들기 1

- Long form으로 되어 있는 종단 자료를 제공해 주기 때문에 각 년도별 자료를 가공할 필요가 없음
- long form 자료를 불러온 후 사용할 변수를 선택하고 재부호화하는 작업을 진행
- 필요한 변수만을 선택
- listwise deletion 후 분석 진행

종단 자료 만들기 2

```
1.  install.packages("foreign")
2.  library(foreign)
3.  setwd("C:\\hyun\\hyun\\2024_spring\\복지패널\\data\\한국복지패널_1_18차_결합데이터_spss/")

4.  pdt <- read.spss("koweps_hp01_18_long_240329.sav", use.value.labels = F, to.data.frame=T)

5.  table(pdt$p02_4aq1, useNA="ifany") ## union
6.  pdt$union <- ifelse(pdt$p02_4aq1==1,0,
7.                  ifelse(pdt$p02_4aq1==2 | pdt$p02_4aq1==3, 1,
8.                         ifelse(pdt$p02_4aq1==4,2,NA)))
9.  table(pdt$p02_4aq1, pdt$union, useNA="ifany")

10. table(pdt$h_g8, pdt$h_g9, useNA="ifany") ## 장애종류 및 등급
11. pdt$disa <- ifelse(pdt$h_g8>=1 & pdt$h_g8<=16,1,
12.                    ifelse(pdt$h_g8==0,0,NA))
13. table(pdt$h_g8, pdt$disa, useNA="ifany")

14. table(pdt$p02_8, pdt$p02_1, useNA="ifany")
15. pdt$wtime <- ifelse(pdt$p02_8<=154, pdt$p02_8, NA)
16. table(pdt$wtime, useNA="ifany")
17. pdt$wtime <- pdt$wtime*4.345   # 한달은 4.345 주
18. table(pdt$p02_8aq1, useNA="ifany") #한달 평균 임금 8-17차 조사
19. summary(pdt$p02_8aq1)#시간당 임금

20. pdt <- pdt[pdt$wtime!=0 & !is.na(pdt$wtime),]

21. pdt$hwage <- pdt$p02_8aq1*10000/pdt$wtime

22. pdt <- pdt[pdt$hwage>0 & !is.na(pdt$hwage),]

23. summary(pdt$hwage)
24. pdt$lhwage <- log(pdt$hwage)summary(pdt$lhwage)
25. hist(pdt$lhwage)
```

종단 자료 만들기 3

```
### confounding
table(pdt$year, useNA="ifany") # 응답년도
table(pdt$h_g4, useNA="ifany")
pdt$age <- pdt$year-pdt$h_g4  # 연령
table(pdt$age, useNA="ifany")
plot(pdt$age, pdt$lhwage)

table(pdt$h_g3,useNA="ifany")  # 성별
pdt$gend <- pdt$h_g3-1

table(pdt$h_g6, useNA="ifany") # 교육수준
pdt$edu <- c(0,0,1,2,3,4,5,5)[match(pdt$h_g6, 2:9)]
table(pdt$h_g6, pdt$edu, useNA="ifany")

table(pdt$h_g10, useNA="ifany") # 혼인상태
pdt$marst <- c(NA,1,2,3,3,0)[match(pdt$h_g10, 0:5)]
table(pdt$h_g10, pdt$marst, useNA="ifany")  # 미혼, 유배우자, 사별, 이혼### 일자리 변수

table(pdt$p02_4, useNA="ifany") ## 근무 시작 년, 15차 없음
pdt$jdur <- pdt$year-pdt$p02_4
table(pdt$jdur, useNA="ifany")

table(pdt$h_reg5,useNA="ifany")  # 5개 권역별 지역구분
pdt$reg <- pdt$h_reg5-1 #서울, 광역시, 시, 군, 도농복합군

adt <- subset(pdt, select=c(h_pid, wv, union, hwage, lhwage, disa,
                            year, age, gend, edu, marst, jdur, reg))
adt <- adt[complete.cases(adt),]

write.csv(adt, "C:\\hyun\\hyun\\2024_spring\\복지패널\\data\\adt.csv")
save(adt, file="C:\\hyun\\hyun\\2024_spring\\복지패널\\data\\adt.RData")
```

2: 기초 분석

- 기초 분석

기초 분석 1

- 평균과 다섯 요약값 보기
  - summary(adt)
    - 논리적으로 가능하지 않은 값은 없는가?
    - 결측값은 없는가?
    - 변수들의 분포는 적절한가?

```
> summary(adt)
     h_pid              wv            union           hwage
 Min.   :     301   Min.   : 8.00   Min.   :0.0000   Min.   :    27.6
 1st Qu.: 222901   1st Qu.:10.00   1st Qu.:0.0000   1st Qu.:  7767.5
 Median : 422301   Median :13.00   Median :0.0000   Median : 11737.6
 Mean   : 469726   Mean   :12.96   Mean   :0.3548   Mean   : 15144.2
 3rd Qu.: 669551   3rd Qu.:16.00   3rd Qu.:1.0000   3rd Qu.: 18642.1
 Max.   :1201001   Max.   :18.00   Max.   :2.0000   Max.   :598389.0
     lhwage            disa             year          age             gend
 Min.   : 3.318   Min.   :0.00000   Min.   :2012   Min.   :18.00   Min.   :0.0000
 1st Qu.: 8.958   1st Qu.:0.00000   1st Qu.:2014   1st Qu.:35.00   1st Qu.:0.0000
 Median : 9.371   Median :0.00000   Median :2017   Median :45.00   Median :0.0000
 Mean   : 9.409   Mean   :0.04737   Mean   :2017   Mean   :46.55   Mean   :0.4843
 3rd Qu.: 9.833   3rd Qu.:0.00000   3rd Qu.:2020   3rd Qu.:56.00   3rd Qu.:1.0000
 Max.   :13.302   Max.   :1.00000   Max.   :2022   Max.   :93.00   Max.   :1.0000
      edu             marst             jdur            reg
 Min.   :0.000   Min.   :0.0000   Min.   : 0.00   Min.   :0.000
 1st Qu.:2.000   1st Qu.:1.0000   1st Qu.: 1.00   1st Qu.:1.000
 Median :2.000   Median :1.0000   Median : 3.00   Median :2.000
 Mean   :2.579   Mean   :0.9689   Mean   : 5.97   Mean   :1.511
 3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.: 9.00   3rd Qu.:2.000
 Max.   :5.000   Max.   :3.0000   Max.   :61.00   Max.   :4.000
```

기초 분석 2

```
1.  dsc <- table(adt$union)
2.  dsc <- c(dsc, prop.table(table(adt$union)))
3.  dsc <- c(dim(adt)[1],-100,dsc[1],-dsc[4],dsc[2],-dsc[5],dsc[3],-dsc[6])

4.  vcat <- c(0,0,1,1,0,1,1,1,0,1)

5.  for (i in 4:13) {
6.  vn <- names(adt)[i]
7.  if (vcat[i-3]==1) {
8.     eval(parse(text=paste("tb <- table(adt$",vn,")", sep="")))
9.     eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
10.                    vn, "))*(-100))", sep="")))
11.     eval(parse(text=paste("tb <- cbind(tb, table(adt$",
12.                    vn,"[adt$union==0]))", sep="")))
13.    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
14.                    vn, "[adt$union==0]))*(-100))", sep="")))
15.    eval(parse(text=paste("tb <- cbind(tb, table(adt$",
16.                    vn, "[adt$union==1]))", sep="")))
17.    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
18.                    vn, "[adt$union==1]))*(-100))", sep="")))
19.    eval(parse(text=paste("tb <- cbind(tb, table(adt$",
20.                    vn, "[adt$union==2]))", sep="")))
21.    eval(parse(text=paste("tb <- cbind(tb, prop.table(table(adt$",
22.                    vn, "[adt$union==2]))*(-100))", sep="")))
23.     }
```

기초 분석 3

```
1.      else  eval(parse(text=paste("tb <- c(mean(adt$",
2.                                  vn,"),-sd(adt$",vn,"), mean(adt$",
3.                                 vn,"[adt$union==0]),-sd(adt$",
4.                                  vn, "[adt$union==0]), mean(adt$",
5.                                 vn,"[adt$union==1]),-sd(adt$",
6.                                 vn, "[adt$union==1]), mean(adt$ " ,
7.                                 vn,"[adt$union==2]),-sd(adt$",
8.                                 vn,"[adt$union==2]))",
9.                                  sep="")))
10.      dsc <- rbind(dsc, tb)
11.     }

12.     write.csv(dsc, "C:\\hyun\\hyun\\2024_spring\\복지패널\\data\\dsc.csv")
```

기초 분석 4

	전체		무조합원		비조합원		조합원	
	N/M	P/SD	N/M	P/SD	N/M	P/SD	N/M	P/SD
전체	47,645	(100.0)	35,306	(0.7)	7,772	(0.2)	4,567	(0.1)
시간임금	15,144.2	(12,090.4)	13,394.3	(10,516.4)	17,643.0	(14,886.6)	24,419.4	(13,139.0)
로그임금	9.4	(0.6)	9.3	(0.6)	9.5	(0.7)	10.0	(0.5)
장애인여부								
비장애인	45,388	(95.3)	33,747	(95.6)	7,218	(92.9)	4,423	(96.8)
장애인	2,257	(4.7)	1,559	(4.4)	554	(7.1)	144	(3.2)
응답년도								
2012	4,738	(9.9)	3,642	(10.3)	600	(7.7)	496	(10.9)
2013	4,677	(9.8)	3,658	(10.4)	607	(7.8)	412	(9.0)
2014	4,620	(9.7)	3,489	(9.9)	684	(8.8)	447	(9.8)
2015	4,528	(9.5)	3,424	(9.7)	654	(8.4)	450	(9.9)
2016	4,491	(9.4)	3,398	(9.6)	649	(8.4)	444	(9.7)
2017	4,553	(9.6)	3,359	(9.5)	782	(10.1)	412	(9.0)
2018	4,525	(9.5)	3,309	(9.4)	760	(9.8)	456	(10.0)
2020	4,291	(9.0)	3,073	(8.7)	764	(9.8)	454	(9.9)
2021	5,611	(11.8)	4,048	(11.5)	1,070	(13.8)	493	(10.8)
2022	5,611	(11.8)	3,906	(11.1)	1,202	(15.5)	503	(11.0)
연령	46.6	(15.1)	45.9	(14.8)	51.1	(17.8)	43.7	(9.9)
성별								
남성	24,569	(51.6)	17,521	(49.6)	3,841	(49.4)	3,207	(70.2)
여성	23,076	(48.4)	17,785	(50.4)	3,931	(50.6)	1,360	(29.8)
교육수준								
초졸이하	5,004	(10.5)	3,328	(9.4)	1,588	(20.4)	88	(1.9)
중졸	3,860	(8.1)	3,163	(9.0)	496	(6.4)	201	(4.4)
고졸	15,249	(32.0)	12,295	(34.8)	1,453	(18.7)	1,501	(32.9)
전문대졸	7,767	(16.3)	6,117	(17.3)	950	(12.2)	700	(15.3)
대졸	13,592	(28.5)	9,226	(26.1)	2,585	(33.3)	1,781	(39.0)
대학원이상	2,173	(4.6)	1,177	(3.3)	700	(9.0)	296	(6.5)
혼인상태								
미혼	10,988	(23.1)	8,726	(24.7)	1,457	(18.7)	805	(17.6)
유배우자	30,343	(63.7)	21,972	(62.2)	4,869	(62.6)	3,502	(76.7)
사별	3,123	(6.6)	1,999	(5.7)	1,059	(13.6)	65	(1.4)
이혼	3,191	(6.7)	2,609	(7.4)	387	(5.0)	195	(4.3)
근속년수	6.0	(7.4)	4.8	(6.2)	7.4	(8.8)	12.7	(9.4)
거주지역								
서울특별시	8,260	(17.3)	6,582	(18.6)	1,070	(13.8)	608	(13.3)
광역시	13,531	(28.4)	10,141	(28.7)	2,110	(27.1)	1,280	(28.0)
시	20,328	(42.7)	14,920	(42.3)	3,281	(42.2)	2,127	(46.6)
군	4,303	(9.0)	2,844	(8.1)	1,064	(13.7)	395	(8.6)
도농복합군	1,223	(2.6)	819	(2.3)	247	(3.2)	157	(3.4)

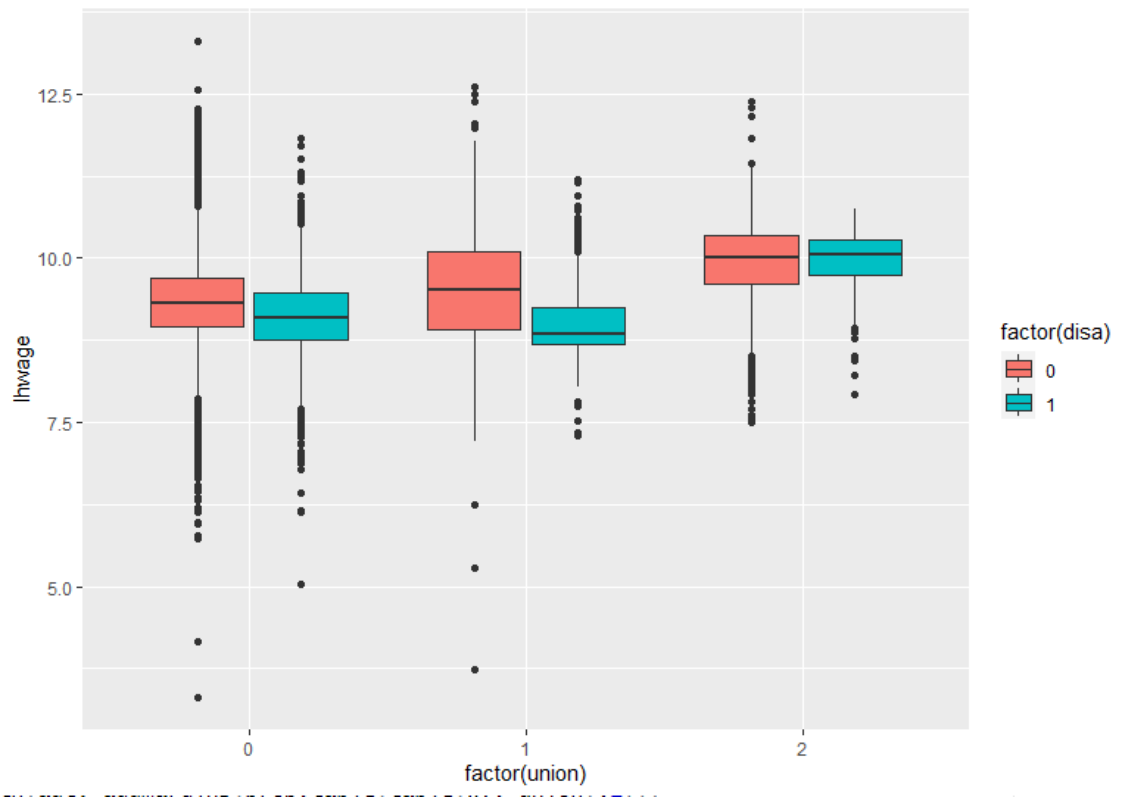
기초 분석 5

```
> with(adt, addmargins(table(wv, union)))
     union
wv       0     1     2   Sum
  8   3642   600   496  4738
  9   3658   607   412  4677
  10  3489   684   447  4620
  11  3424   654   450  4528
  12  3398   649   444  4491
  13  3359   782   412  4553
  14  3309   760   456  4525
  16  3073   764   454  4291
  17  4048  1070   493  5611
  18  3906  1202   503  5611
  Sum 35306  7772  4567 47645
> with(adt, addmargins(prop.table(table(wv, union),1)))
     union
wv             0          1          2        Sum
  8   0.76867877 0.12663571 0.10468552 1.00000000
  9   0.78212529 0.12978405 0.08809066 1.00000000
  10  0.75519481 0.14805195 0.09675325 1.00000000
  11  0.75618375 0.14443463 0.09938163 1.00000000
  12  0.75662436 0.14451124 0.09886440 1.00000000
  13  0.73775533 0.17175489 0.09048979 1.00000000
  14  0.73127072 0.16795580 0.10077348 1.00000000
  16  0.71615008 0.17804708 0.10580284 1.00000000
  17  0.72144003 0.19069685 0.08786313 1.00000000
  18  0.69613260 0.21422206 0.08964534 1.00000000
  Sum 7.42155572 1.61609426 0.96235002 10.00000000
```

- 연도별 노동조합 변수 변화 보기
  - with(adt, addmargins(table(wv, union)))
  - with(adt, addmargins(prop.table(table(wv, union),1)))
  - 17차년도 사례수 증가
  - 무조합원 감소, 비조합원 증가, 조합원 유사한 수준 유지
  - 17, 18차는 2021년과 2022년으로 코로나 상황

기초 분석 6

- 임금의 분포
  - library(ggplot2)
  - ggplot(adt, aes(x=factor(union), y=lhwage, fill=factor(disa))) +
  - geom_boxplot()



기초 분석 7

- 장애유무에 따른 노동조합 임금프리미엄
- tapply(adt$lhwage, adt$union, mean)
- summary(aov(lhwage ~ factor(union), data=adt))

- tapply(adt$lhwage[adt$disa==0], adt$union[adt$disa==0], mean)
- summary(aov(lhwage ~ factor(union), data=adt[adt$disa==0,]))

- tapply(adt$lhwage[adt$disa==1], adt$union[adt$disa==1], mean)
- summary(aov(lhwage ~ factor(union), data=adt[adt$disa==1,]))

```
> tapply(adt$lhwage, adt$union, mean)
       0        1        2
9.314926 9.507204 9.969729
> summary(aov(lhwage ~ factor(union), data=adt))
                Df Sum Sq Mean Sq F value Pr(>F)
factor(union)    2   1823   911.7    2375 <2e-16 ***
Residuals    47642  18292     0.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> tapply(adt$lhwage[adt$disa==0], adt$union[adt$disa==0], mean)
       0        1        2
9.324153 9.544386 9.969617
> summary(aov(lhwage ~ factor(union), data=adt[adt$disa==0,]))
                Df Sum Sq Mean Sq F value Pr(>F)
factor(union)    2   1758   878.8    2328 <2e-16 ***
Residuals    45385  17131     0.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> tapply(adt$lhwage[adt$disa==1], adt$union[adt$disa==1], mean)
       0        1        2
9.115207 9.022767 9.973165
> summary(aov(lhwage ~ factor(union), data=adt[adt$disa==1,]))
                Df Sum Sq Mean Sq F value Pr(>F)
factor(union)    2  108.4   54.21   127.8 <2e-16 ***
Residuals     2254  955.7    0.42
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3: 일반최소제곱법

- 일반최소제곱법
- Sandwich 추정법

일반최소제곱법 1

- 어떤 특정 변수의 값을 예측하고자 할 때 가장 많이 사용하는 추정법이 ordinary least squares (OLS:일반최소제곱법)이다
- 우리의 모형을 다음과 같이 나타낸다고 하자
- $y_i = \sum_k \beta_k X_{ki} + \epsilon_i$ 혹은 $Y = X^T\beta + \epsilon$

  where K는 모수의 개수(변수 수+1)이며 i는 개별 사례를 나타낸다
- OLS는 β를 다음과 같이 추정한다
- $\min_{\beta} \epsilon'\epsilon = \min_{\beta}[Y - X^T\beta]^T[Y - X^T\beta]$
  - $\hat{\beta} = (X^TX)^{-1}X^TY$
  - $\widehat{var}(\hat{\beta}) = var[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T var(\epsilon)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \frac{1}{N-K}(\hat{\epsilon}^T\hat{\epsilon})(X^TX)^{-1}$

일반최소제곱법 2

- Gauss-Markov Theorem에 따르면 다음의 조건 하에 위의 추정치가 BLUE(best linear unbiased estimator)라는 것을 보여준다
  - best는 가장 효율적이라는 의미이다
- 조건 1: $E(\epsilon_i) = 0$, 모든 i에 대해
- 조건 2: $var(\epsilon_i) = \sigma^2$, 모든 i에 대해
  - 동분산성(homoskedasticity) 가정
- 조건 3: $cov(\epsilon_i, \epsilon_j) = 0$, 모든 $i \neq j$에 대해
  - 잔차의 독립성(independence) 가정
- 조건 4: $cov(x_{ki}, \epsilon_i) = 0$, 모든 k 및 i에 대해
  - 외생성(exogeneity) 가정

일반최소제곱법 3

- OLS에서 어떤 계수 $\hat{\beta}_k$의 해석
  - 다른 변수를 통제한 상태에서 X_k가 한 단위 증가할 때 Y의 변화량
- 특정 모형의 예측력이 얼마나 좋은가를 평가하는 하나의 통계치로 때때로 다중결정계수(coefficient of multiple determination)라 불리는 R^2가 쓰인다
  - $R^2 = \frac{TSS-SSE}{TSS} = \frac{\sum(Y-\bar{Y})^2-\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = corr(Y,\hat{Y})^2$
- t-test:
  - $H_0: \beta_k = 0, H_A: \beta_k \neq 0$
  - $t = \frac{\hat{\beta}_k}{se}$를 사용하며 이 통계치는 $df = N - K$인 t-분포를 따른다
- F-test
  - H_0: $\beta_2 = \cdots = \beta_K = 0, (\beta_1\ is\ for\ the\ intercept), H_A$: at least one β is not zero
  - $F = \frac{R^2/(K-1)}{(1-R^2)/[N-K]}$ 값은 $df_1 = K - 1$ 이고 $df_2 = N - K$인 F-분포를 따른다

일반최소제곱법 4

- $\ln(hwage) = \beta_0 + \beta_1 Union1 + \beta_2 Union2 + X^T\alpha + \epsilon$
  - adt$agesq <- adt$age^2
  - fit1 <- lm(lhwage ~ factor(union) +factor(disa) + factor(year) + age + agesq + factor(gend) + factor(marst) + jdur+factor(reg), data=adt)
  - summary(fit1)
- 무노조에 비해 비노조인 경우 시간당 임금이 exp(0.198) ≈ 1.219 배 이다
- 무노조에 비해 노조원인 경우 시간당 임금이 exp(0.277) ≈ 1.319 배 이다

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.826e+00	2.417e-02	365.189	< 2e-16	***
factor(union)1	1.983e-01	6.062e-03	32.717	< 2e-16	***
factor(union)2	2.772e-01	7.802e-03	35.532	< 2e-16	***
factor(disa)1	-1.184e-01	1.030e-02	-11.493	< 2e-16	***
factor(year)2013	6.714e-02	9.654e-03	6.955	3.56e-12	***
factor(year)2014	7.707e-02	9.686e-03	7.957	1.80e-15	***
factor(year)2015	1.444e-01	9.737e-03	14.827	< 2e-16	***
factor(year)2016	1.917e-01	9.763e-03	19.632	< 2e-16	***
factor(year)2017	2.658e-01	9.741e-03	27.281	< 2e-16	***
factor(year)2018	3.238e-01	9.765e-03	33.160	< 2e-16	***
factor(year)2020	4.381e-01	9.922e-03	44.151	< 2e-16	***
factor(year)2021	4.843e-01	9.340e-03	51.849	< 2e-16	***
factor(year)2022	5.177e-01	9.358e-03	55.323	< 2e-16	***
age	2.512e-02	1.049e-03	23.959	< 2e-16	***
agesq	-4.157e-04	1.007e-05	-41.286	< 2e-16	***
factor(gend)1	-2.769e-01	4.537e-03	-61.023	< 2e-16	***
factor(marst)1	1.977e-01	7.189e-03	27.498	< 2e-16	***
factor(marst)2	1.959e-01	1.239e-02	15.805	< 2e-16	***
factor(marst)3	2.786e-02	1.110e-02	2.509	0.01211	*
jdur	2.853e-02	3.306e-04	86.300	< 2e-16	***
factor(reg)1	-1.155e-01	6.563e-03	-17.595	< 2e-16	***
factor(reg)2	-7.539e-02	6.162e-03	-12.234	< 2e-16	***
factor(reg)3	-1.625e-01	8.925e-03	-18.207	< 2e-16	***
factor(reg)4	-5.035e-02	1.440e-02	-3.498	0.00047	***

Sandwich 추정법 I

- 앞서 $var(\epsilon) = \sigma^2$로 가정하였으며 이는 조건 2에 의한 것이다.
- $\widehat{var}(\hat{\beta}) = (X^T X)^{-1} X^T var(\epsilon) X (X^T X)^{-1}$
- 하지만 패널 자료는 한 사례당 여러 년도의 관측값이 있기 때문에 사례에 따라 다른 분산을 가지고 있을 가능성이 높다
  - 어떤 패널들은 월급의 변동이 클 것이고, 어떤 패널들은 월급의 변동이 적을 것이다
- 잔차 가정이 만족스럽지 못하다면 하나의 대안은 Sandwich 추정법으로 분산을 구하는 것이다
  - 흔히 $(X^T X)^{-1}$를 bread part라고 하고 $X^T var(\epsilon) X$를 meat part라고 한다
  - 또한 많은 경우 $var(\epsilon)$를 Ω로 표시한다

Sandwich 추정법 2

- 모든 잔차 사이에 이분산성(heteroskedasticity)이 있으면
  - $X^T \widehat{var}(\epsilon) X$에서 $\widehat{var}(\epsilon) = \hat{\Omega} = \frac{N}{N-K} \mathrm{diag}(\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_N)$ 으로 추정하고 $\hat{\epsilon}$는 추정된 잔차임
- 만약 집단 내 상관관계를 고려하고 싶다면
  - $X^T \widehat{var}(\epsilon) X = \sum_{g=1}^{G} \hat{u}_g^T \hat{u}_j \, , \hat{u}_j = \sum_{i=1}^{N_G} \hat{\epsilon}_i * x_i$
  - 표본의 수가 무한대일 수 없다는 점을 교정하기 위해 (finite-sample adjustment) $(N-1)G/[(N-K)(G-1)]$를 곱해준다

Sandwich 추정법 3

- 통상의 추정법
  - fit1 <- lm(lhwage ~ factor(union) +factor(disa) +
  - factor(year) + age + agesq + factor(gend) +
  - factor(marst) + jdur+factor(reg),
  - data=adt)
- Sandwich 추정법
  - #install.packages(sandwich)
  - library(sandwich)
  - library(lmtest)
  - coeftest(fit1, df=Inf, vcov=vcovHC(fit1, type= " HC1 "))
- Clustered sandwich 추정법
  - #install.packages("multiwayvcov")
  - library(multiwayvcov)
  - fit1_clv <- cluster.vcov(fit1, adt$pid)
  - coeftest(fit1, vcov = fit1_clv)

Sandwich 추정법 4

OLS

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.826e+00	2.417e-02	365.189	< 2e-16	***
factor(union)1	1.983e-01	6.062e-03	32.717	< 2e-16	***
factor(union)2	2.772e-01	7.802e-03	35.532	< 2e-16	***
factor(disa)1	-1.184e-01	1.030e-02	-11.493	< 2e-16	***
factor(year)2013	6.714e-02	9.654e-03	6.955	3.56e-12	***
factor(year)2014	7.707e-02	9.686e-03	7.957	1.80e-15	***
factor(year)2015	1.444e-01	9.737e-03	14.827	< 2e-16	***
factor(year)2016	1.917e-01	9.763e-03	19.632	< 2e-16	***
factor(year)2017	2.658e-01	9.741e-03	27.281	< 2e-16	***
factor(year)2018	3.238e-01	9.765e-03	33.160	< 2e-16	***
factor(year)2020	4.381e-01	9.922e-03	44.151	< 2e-16	***
factor(year)2021	4.843e-01	9.340e-03	51.849	< 2e-16	***
factor(year)2022	5.177e-01	9.358e-03	55.323	< 2e-16	***
age	2.512e-02	1.049e-03	23.959	< 2e-16	***
agesq	-4.157e-04	1.007e-05	-41.286	< 2e-16	***
factor(gend)1	-2.769e-01	4.537e-03	-61.023	< 2e-16	***
factor(marst)1	1.977e-01	7.189e-03	27.498	< 2e-16	***
factor(marst)2	1.959e-01	1.239e-02	15.805	< 2e-16	***
factor(marst)3	2.786e-02	1.110e-02	2.509	0.01211	*
jdur	2.853e-02	3.306e-04	86.300	< 2e-16	***
factor(reg)1	-1.155e-01	6.563e-03	-17.595	< 2e-16	***
factor(reg)2	-7.539e-02	6.162e-03	-12.234	< 2e-16	***
factor(reg)3	-1.625e-01	8.925e-03	-18.207	< 2e-16	***
factor(reg)4	-5.035e-02	1.440e-02	-3.498	0.00047	***

Sandwich

	Estimate	Std. Error	z value	Pr(>|z|)	
ercept)	8.8260e+00	2.5279e-02	349.1468	< 2.2e-16	***
or(union)1	1.9833e-01	6.1909e-03	32.0363	< 2.2e-16	***
or(union)2	2.7722e-01	7.1289e-03	38.8872	< 2.2e-16	***
or(disa)1	-1.1842e-01	1.1763e-02	-10.0677	< 2.2e-16	***
or(year)2013	6.7145e-02	1.0608e-02	6.3294	2.461e-10	***
or(year)2014	7.7072e-02	1.0327e-02	7.4634	8.431e-14	***
or(year)2015	1.4438e-01	1.0277e-02	14.0492	< 2.2e-16	***
or(year)2016	1.9167e-01	1.0185e-02	18.8195	< 2.2e-16	***
or(year)2017	2.6575e-01	1.0160e-02	26.1564	< 2.2e-16	***
or(year)2018	3.2380e-01	9.9836e-03	32.4328	< 2.2e-16	***
or(year)2020	4.3807e-01	9.8208e-03	44.6060	< 2.2e-16	***
or(year)2021	4.8429e-01	9.3762e-03	51.6513	< 2.2e-16	***
or(year)2022	5.1773e-01	9.4727e-03	54.6553	< 2.2e-16	***
	2.5123e-02	1.1090e-03	22.6531	< 2.2e-16	***
q	-4.1572e-04	1.0966e-05	-37.9085	< 2.2e-16	***
or(gend)1	-2.7685e-01	4.5305e-03	-61.1084	< 2.2e-16	***
or(marst)1	1.9767e-01	6.9983e-03	28.2458	< 2.2e-16	***
or(marst)2	1.9588e-01	1.2965e-02	15.1085	< 2.2e-16	***
or(marst)3	2.7856e-02	1.1523e-02	2.4174	0.01563	*
	2.8531e-02	3.7902e-04	75.2753	< 2.2e-16	***
or(reg)1	-1.1547e-01	6.8624e-03	-16.8266	< 2.2e-16	***
or(reg)2	-7.5391e-02	6.4694e-03	-11.6535	< 2.2e-16	***
or(reg)3	-1.6249e-01	9.3965e-03	-17.2929	< 2.2e-16	***
or(reg)4	-5.0350e-02	1.2864e-02	-3.9142	9.071e-05	***

Clustered sandwich

	Estimate	Std. Error	t value	Pr(>|t|)	
	8.8260e+00	4.1625e-02	212.0340	< 2.2e-16	***
	1.9833e-01	9.8556e-03	20.1238	< 2.2e-16	***
	2.7722e-01	1.3345e-02	20.7739	< 2.2e-16	***
	-1.1842e-01	2.1868e-02	-5.4154	6.144e-08	***
	6.7145e-02	7.8278e-03	8.5777	< 2.2e-16	***
	7.7072e-02	7.9569e-03	9.6862	< 2.2e-16	***
	1.4438e-01	8.1370e-03	17.7434	< 2.2e-16	***
	1.9167e-01	8.4391e-03	22.7127	< 2.2e-16	***
	2.6575e-01	8.5810e-03	30.9696	< 2.2e-16	***
	3.2380e-01	8.8304e-03	36.6684	< 2.2e-16	***
	4.3807e-01	9.0787e-03	48.2524	< 2.2e-16	***
	4.8429e-01	9.0939e-03	53.2550	< 2.2e-16	***
	5.1773e-01	9.2215e-03	56.1440	< 2.2e-16	***
	2.5123e-02	1.9283e-03	13.0283	< 2.2e-16	***
	-4.1572e-04	1.8825e-05	-22.0833	< 2.2e-16	***
	-2.7685e-01	8.8624e-03	-31.2389	< 2.2e-16	***
	1.9767e-01	1.4015e-02	14.1049	< 2.2e-16	***
	1.9588e-01	2.3873e-02	8.2051	2.361e-16	***
	2.7856e-02	2.2063e-02	1.2626	0.20675	
	2.8531e-02	6.8673e-04	41.5465	< 2.2e-16	***
	-1.1547e-01	1.3413e-02	-8.6092	< 2.2e-16	***
	-7.5391e-02	1.2606e-02	-5.9804	2.242e-09	***
	-1.6249e-01	1.7732e-02	-9.1640	< 2.2e-16	***
	-5.0350e-02	2.2761e-02	-2.2121	0.02696	*

4: 고정효과 모형

- 기본 모형 및 추정방법
- within 추정
- LSDV 추정
- 1차 차분 추정

기본 모형

- 다음의 패널 선형회귀모형을 가정하자
- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}, \quad i = 1, \ldots, n$ 및 $t = 1, \ldots, T$
- 에서 오차항은 시간에 따라 변하지 않는 패널의 개체 특성을 나타내는 u_i 와 시간과 패널 개체에 따라 변하는 순수한 오차항인 e_{it}로 구성되어 있다
- 이제 오차항 u_i를 확률변수(random variable)가 아닌 추정해야 할 모수(parameter)로 간주하는 고정효과(fixed effects) 모형에 대해 알아보자
  - 다른 말로 하면 u_i가 어떤 특정한 분포를 따르지 않는다는 것이다
- 앞이 식을 달리 쓰면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
  - 고정효과 모형은 상수항이 패널 개체 별로 서로 다르면서 고정되어(fixed) 있다고 가정한다

Within 추정 1

- within 추정법을 알아보기 위해 각 패널의 연도별 변수값에 대해 평균을 구한 between 모형을 적어보자
- $\bar{y}_i = \alpha + \beta \bar{x}_i + u_i + \bar{e}_i$
- 이제 앞선 식에서 between 모형을 빼 주면 (within 변환)
- $(y_{it} - \bar{y}_i) = (\alpha - \alpha) + \beta(x_{it} - \bar{x}_i) + (u_i - u_i) + (e_{it} - \bar{e}_i) = \beta(x_{it} - \bar{x}_i) + (e_{it} - \bar{e}_i)$
- α와 u_i가 사라진 것을 알 수 있다 (differenced out). 따라서 $cov(x_{it}, u_i) \neq 0$ 라고 하더라도 OLS 추정을 통해 β에 대한 일치추정량을 구할 수 있다
- 위 식를 추정한 후 u_i를 사후적으로 계산할 수 있다

Within 추정 2

- install.packages("plm")
- library(plm)
- padt <- pdata.frame(adt, index=c("h_pid","wv"), drop.index=TRUE)
- fit2 <- plm(lhwage ~ factor(union) +factor(disa) + factor(year) +
- age + agesq + factor(gend) + factor(marst) +
- jdur+factor(reg),
- data=padt, model="within")
- summary(fit2)

- age와 gend의 계수가 없다
- 계수 크기가 많이 줄어들었다

```
Coefficients: (1 dropped because of singularities)
                    Estimate  Std. Error  t-value  Pr(>|t|)
factor(union)1    2.5597e-02  5.6620e-03   4.5209 6.176e-06 ***
factor(union)2    6.8102e-02  8.1773e-03   8.3282 < 2.2e-16 ***
factor(disa)1    -3.6206e-02  3.8193e-02  -0.9479   0.34316
factor(year)2013  9.8794e-02  6.7688e-03  14.5955 < 2.2e-16 ***
factor(year)2014  1.5795e-01  7.6475e-03  20.6540 < 2.2e-16 ***
factor(year)2015  2.4859e-01  8.8710e-03  28.0228 < 2.2e-16 ***
factor(year)2016  3.2617e-01  1.0269e-02  31.7640 < 2.2e-16 ***
factor(year)2017  4.3551e-01  1.1851e-02  36.7484 < 2.2e-16 ***
factor(year)2018  5.3051e-01  1.3579e-02  39.0673 < 2.2e-16 ***
factor(year)2020  7.1000e-01  1.7269e-02  41.1140 < 2.2e-16 ***
factor(year)2021  7.9865e-01  1.9107e-02  41.7996 < 2.2e-16 ***
factor(year)2022  8.7685e-01  2.1110e-02  41.5362 < 2.2e-16 ***
agesq            -3.9637e-04  2.0822e-05 -19.0362 < 2.2e-16 ***
factor(marst)1    3.3320e-02  1.3600e-02   2.4500   0.01429 *
factor(marst)2    3.1802e-02  2.5450e-02   1.2496   0.21146
factor(marst)3    4.3418e-02  2.2414e-02   1.9371   0.05275 .
jdur              7.1742e-03  4.2449e-04  16.9007 < 2.2e-16 ***
factor(reg)1      1.9141e-02  1.8608e-02   1.0286   0.30367
factor(reg)2     -1.9629e-03  1.4803e-02  -0.1326   0.89451
factor(reg)3     -3.6974e-02  2.5061e-02  -1.4753   0.14013
factor(reg)4     -4.3316e-02  3.6958e-02  -1.1720   0.24119
```

Within 추정 3

- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$에서 $u_i = \tau z_i$로 볼 수 있다
- 이때 z_i는 시간불변인 패널 개체 별 특성, 예를 들어 IQ나 외모와 같은 것이며 τ는 이와 연관된 계수이다. 이런 면에서 u_i는 관찰되지 않은 시간불변 변수의 *시간불변* 효과를 통제하는 측면이 있다
- 헌데, u_i는 모형 추정과정에서 빠지므로 시간불변 변수의 효과를 추정하지 못한다는 약점이 있다
- 예를 들어, 성별과 같은 패널 내 불변변수의 효과는 추정할 수 없다

LSDV 추정

- 앞선 모형에 따르면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
- u_i를 고정효과로 본다면 $(\alpha + u_i)$는 각 개체가 평균으로부터 얼마나 떨어져 있는가를 나타내는 것으로 볼 수 있다
- 이를 달리 모수화한다면 평균으로부터 떨어져 있는 것이 아닌 각 개체의 상수항을 추정하는 다음과 같은 모형으로 쓸 수 있다
- $y_{it} = \sum_{i=1}^{n} \alpha_i + \beta x_{it} + e_{it}$
- 즉 각 패널의 더미변수를 만들어 이를 OLS로 추정하면 고정효과 모형을 추정할 수 있다는 것이다
- 이러한 추정방법을 LSDV(least squares dummy variable) 추정법이라 한다
- Stata에서는 “areg” 명령어가 있으나 R에서는 없는 것으로 판단된다
- lm 명령어로 h_pid를 factor로 처리하면 되지만 추정에 너무 많은 시간이 걸린다

1차 차분 추정 1

- 이제 세 번의 시기에 걸친 자료가 있다고 하자
- $y_{i1} = \alpha + \beta x_{i1} + u_i + e_{i1}, \quad t = 1$
- $y_{i2} = \alpha + \beta x_{i2} + u_i + e_{i2}, \quad t = 2$
- $y_{i3} = \alpha + \beta x_{i3} + u_i + e_{i3}, \quad t = 3$
- 1차 차분을 실행하면
- $\Delta y_{i2} = \Delta\beta x_{i2} + \Delta e_{i2}, \quad t = 2$
- $\Delta y_{i3} = \Delta\beta x_{i3} + \Delta e_{i3}, \quad t = 3$
- $T = 2$인 경우에는 1차 차분을 실행하면 패널자료가 횡단면자료로 전환되지만 $T > 2$이면 1차 차분을 해도 패널자료 구조를 유지하게 된다
- 1차 차분을 합동한 다음 OLS를 적용하면
  - u_i가 없어져서 $corr(x_{it}, u_i) \neq 0$이더라도 일치추정량을 구할 수 있다
  - 그러나 오차항 e_{it}에 자기상관이 존재하지 않더라도 1차 차분 모형의 오차항 Δe_{it}에는 1계 자기상관이 존재한다. 즉 $cov(\Delta e_{it}, \Delta e_{it-1}) = cov(e_{it} - e_{it-1}, e_{it-1} - e_{it-2}) = -\sigma_e^2$
  - 오히려 e_{it}에 $e_{it} = e_{it-1} + v_{it}$와 같이 1계 자기상관이 있으면, $cov(\Delta e_{it}, \Delta e_{it-1}) = cov(v_{it}, v_{it-1}) = 0$
  - 다시 말해 e_{it}에 계수가 1인 1계 자기상관이 있어야만 효율적 추정량을 얻을 수 있다

1차 차분 추정 2

```
fit3 <- plm(lhwage ~ factor(union) +factor(disa) +
            factor(year) +
            age + agesq + factor(gend) + factor(marst) +
            jdur+factor(reg),
          data=padt, model="fd")
summary(fit3)
```

- e_{it}에 자기상관이 없으면 within 추정은 적절하면 1차 차분 모형은 효율적이지 않다
- 하지만 둘 다 일치추정량이다
- 반대로 e_{it}가 1계 자기상관이 있으면 1차 차분 모형이 적절하며 within 추정이 효율적이지 않다

```
Coefficients: (1 dropped because of singularities)
                 Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)     9.7274e-03  4.3639e-03  2.2290  0.025817 *
factor(union)1  1.6255e-02  5.4533e-03  2.9808  0.002877 **
factor(union)2  4.4676e-02  8.2934e-03  5.3870 7.209e-08 ***
factor(disa)1  -2.9130e-02  5.0695e-02 -0.5746  0.565547
factor(year)2013  9.7230e-02  8.0483e-03 12.0808 < 2.2e-16 ***
factor(year)2014  1.5415e-01  1.3172e-02 11.7027 < 2.2e-16 ***
factor(year)2015  2.4004e-01  1.8070e-02 13.2836 < 2.2e-16 ***
factor(year)2016  3.2169e-01  2.2888e-02 14.0552 < 2.2e-16 ***
factor(year)2017  4.3449e-01  2.7752e-02 15.6561 < 2.2e-16 ***
factor(year)2018  5.2766e-01  3.2770e-02 16.1018 < 2.2e-16 ***
factor(year)2020  7.1218e-01  4.1391e-02 17.2063 < 2.2e-16 ***
factor(year)2021  7.9756e-01  4.6921e-02 16.9981 < 2.2e-16 ***
factor(year)2022  8.7388e-01  5.2927e-02 16.5112 < 2.2e-16 ***
agesq          -4.1719e-04  4.4146e-05 -9.4502 < 2.2e-16 ***
factor(marst)1  7.9579e-04  1.8993e-02  0.0419  0.966579
factor(marst)2 -9.8056e-02  3.5673e-02 -2.7487  0.005986 **
factor(marst)3  6.9168e-03  3.0337e-02  0.2280  0.819651
jdur            4.8592e-03  4.9524e-04  9.8118 < 2.2e-16 ***
factor(reg)1    2.4239e-02  2.3215e-02  1.0441  0.296458
factor(reg)2    6.8883e-03  1.9719e-02  0.3493  0.726853
factor(reg)3   -1.5641e-02  3.0748e-02 -0.5087  0.610980
factor(reg)4    2.1574e-02  4.7715e-02  0.4521  0.651172
```

1차 차분 추정 3

- 1차 차분 모형에서 가장 중요한 가정은 e_{it}에 자기상관이 존재하는가의 여부이므로 이를 검정할 필요가 있다
- [pwfdtest] 명령어는 Wooldridge가 제안한 방법을 명령어로 만들어 놓았다
- e_{it}에 자기상관이 없다는 영가설 하에 $corr(\Delta e_{it}, \Delta e_{it-1}) = \frac{cov(\Delta e_{it}, \Delta e_{it-1})}{\sqrt{var(\Delta e_{it})}\sqrt{var(\Delta e_{it-1})}} = \frac{cov(e_{it}-e_{it-1}, e_{it-1}-e_{it-2})}{\sqrt{var(e_{it}-e_{it-1})}\sqrt{var(e_{it-1}-e_{it-2})}} = \frac{-\sigma_e^2}{2\sigma_e^2} = -0.5$
- 따라서 [pwfdtest]sms 1차 차분 모형 오차항의 1계 자기상관계수가 -0.5인지를 검정한다
- pwfdtest(fit3)
- 검정결과가 유의미하므로 영가설을 기각한다.
- 즉 e_{it} 에 자기상관이 존재한다.

```
> pwfdtest(fit3)

        Wooldridge's first-difference test for serial correlation in panels

data:  fit3
F = 1539, df1 = 1, df2 = 28305, p-value < 2.2e-16
alternative hypothesis: serial correlation in differenced errors
```

5: 확률효과모형

- 기본 모형
- 모형 추정

기본 모형 1

- 여전히 기본 모형은 다음과 같다
- $y_{it} = \alpha + \beta x_{it} + u_i + e_{it}$
- 앞선 장의 고정효과 모형에서는 u_i를 추정해야할 모수(parameter)로 간주하였다
- 이에 반해 u_i를 확률변수(random variable)로 가정하는 것을 확률효과(random effects) 모형이라 한다
- 확률효과 모형에서 오차항들은 일반적으로 $u_i \sim N(0, \sigma_u^2)$과 $e_{it} \sim N(0, \sigma_e^2)$인 것으로 가정한다
- 모형을 다시 쓰면 $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$ 로 표현되며 $(\alpha + u_i)$는 확률변수로 간주되며 $E(\alpha + u_i) = \alpha$이다
- 달리 말해, α는 패널 개체별 상수항의 평균을 뜻하게 된다

기본 모형 2

- 이를 OLS로 추정하는 것은 1계 자기상관으로 인해 효율성 문제가 있다
  - $cov(u_i + e_{it}, u_i + e_{it-1}) = cov(u_i, u_i) + cov(u_i, e_{it-1}) + cov(e_{it}, u_i) + cov(e_{it}, e_{it-1}) = \sigma_u^2 \neq 0$
- OLS 추정을 위해서는 $cov(x_{it}, u_i) = 0$이어야 하는데 이것이 성립한다면 다음과 같은 모형을 추정하는 것이 일치추정량이면서 효율적인 추정량이라는 것이 알려져 있다
- $(y_{it} - \theta_i \bar{y}_i) = \alpha(1 - \theta_i) + \beta(x_{it} - \theta_i \bar{x}_i) + [u_i(1 - \theta_i) + (e_{it} - \theta_i \bar{e}_i)]$
- 여기에서 $\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_u^2 + \sigma_e^2}}$
- 만약 $T_i = T$인 균형패널이면 $\theta_i = \theta$가 된다
- 먼저 $\hat{\sigma}_u^2$와 $\hat{\sigma}_e^2$를 추정한 다음, $\hat{\theta}_i$를 구한 후, 그것을 앞 식에 대입하여 OLS로 추정하는 단계를 거친다

기본 모형 3

- 그런데 고정효과 모형의 오차항 분산은$\sigma_{FE}^2 = var(u_i + e_{it}) = var(e_{it}) = \sigma_e^2$
- between 모형의 오차항 분산은 $\sigma_{BE}^2 = var(u_i + \bar{e}_i) = var(u_i) + var\left(\frac{\sum_{t=1}^{T_i} e_{it}}{T_i}\right) = \sigma_u^2 + \sigma_e^2/T_i$. 즉 $T_i\sigma_u^2 + \sigma_e^2 = T_i\sigma_{BE}^2$
- 따라서 θ_i는 다음과 같이 쓸 수 있다
- $\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i\sigma_u^2+\sigma_e^2}} = 1 - \sqrt{\frac{\sigma_{FE}^2}{T_i\sigma_{BE}^2}}$
- 확률효과 모형 추정량은 결국 between 모형의 추정량과 고정효과 모형의 추정량의 가중평균치라고 할 수 있다

기본 모형 4

- $\theta_i = 0$(즉 $\sigma_u^2 = 0$)이면 패널개체 별 이질성이 없다는 뜻이므로 합동 OLS로 추정하는 것과 같다
- 반면 $\theta_i = 1$(즉 $\sigma_e^2 = 0$)이면 오직 개체별 이질성만이 남기 때문에 u_i가 중요해지고 within 회귀모형과 같은 형태가 된다
- 확률효과 모형의 추정량은 between 모형 추정량보다 효율적이다
  - between 모형은 패널 별 평균만을 변수로 추정하는데 반해 확률효과 모형은 패널 간 정보와 패널 내 정보를 모두 사용한다
- $cov(x_{it}, u_i) = 0$이라는 가정이 성립하면
  - 고정효과 모형은 패널 개체 더미변수를 추정하기 때문에 자유도의 손실이 발생한다
  - 따라서 확률효과 모형의 추정량이 더 효율적이다
  - 또한 $\theta_i \neq 1$이면 시간에 따라 변하지 않는 설명변수(time invariant variable)에 대한 β 추정치도 얻을 수 있는 장점이 있다
- 하지만 $cov(x_{it}, u_i) = 0$이라는 가정이 성립하지 않는다면 확률효과 모형 추정량은 일치추정량이 되지 못한다

모형 추정

- fit4 <- plm(lhwage ~ factor(union) +factor(disa) + factor(year) +
- age + agesq + factor(gend) + factor(marst) +
- jdur+factor(reg),
- data=padt, model="random")
- summary(fit4)
- age와 gend의 계수가 있다

	Estimate	Std. Error	z-value	Pr(>|z|)	
(Intercept)	8.6418e+00	3.0765e-02	280.8987	< 2.2e-16	***
factor(union)1	7.9067e-02	5.3526e-03	14.7717	< 2.2e-16	***
factor(union)2	1.5380e-01	7.6505e-03	20.1028	< 2.2e-16	***
factor(disa)1	-1.1571e-01	1.6480e-02	-7.0213	2.198e-12	***
factor(year)2013	6.7455e-02	6.5726e-03	10.2631	< 2.2e-16	***
factor(year)2014	9.0938e-02	6.6737e-03	13.6264	< 2.2e-16	***
factor(year)2015	1.5348e-01	6.7720e-03	22.6639	< 2.2e-16	***
factor(year)2016	1.9956e-01	6.8492e-03	29.1370	< 2.2e-16	***
factor(year)2017	2.7732e-01	6.9116e-03	40.1246	< 2.2e-16	***
factor(year)2018	3.3915e-01	7.0068e-03	48.4032	< 2.2e-16	***
factor(year)2020	4.5529e-01	7.2912e-03	62.4442	< 2.2e-16	***
factor(year)2021	5.1109e-01	7.1339e-03	71.6423	< 2.2e-16	***
factor(year)2022	5.5223e-01	7.2403e-03	76.2721	< 2.2e-16	***
age	3.2803e-02	1.3516e-03	24.2691	< 2.2e-16	***
agesq	-4.6048e-04	1.2982e-05	-35.4709	< 2.2e-16	***
factor(gend)1	-2.9105e-01	8.0414e-03	-36.1945	< 2.2e-16	***
factor(marst)1	1.4325e-01	9.7331e-03	14.7177	< 2.2e-16	***
factor(marst)2	1.3715e-01	1.6987e-02	8.0737	6.819e-16	***
factor(marst)3	2.7683e-02	1.5380e-02	1.8000	0.07187	.
jdur	1.5897e-02	3.6530e-04	43.5184	< 2.2e-16	***
factor(reg)1	-7.1991e-02	1.0331e-02	-6.9685	3.203e-12	***
factor(reg)2	-3.9134e-02	9.2711e-03	-4.2210	2.432e-05	***
factor(reg)3	-9.7691e-02	1.3752e-02	-7.1040	1.212e-12	***
factor(reg)4	-4.5359e-02	2.1025e-02	-2.1574	0.03097	*

6: Hausman 검정

- 하우스만 검정

하우스만 검정 1

- 패널 모형은 다음과 같이 썼다
- $y_{it} = (\alpha + u_i) + \beta x_{it} + e_{it}$
- 지금까지 배운 것처럼 고정효과 모형에서는 $\alpha + u_i$를 패널 개체별로 고정되어 있는 모수로 해석한다
- 이에 반해 확률효과 모형에서는 $\alpha + u_i$를 확률분포를 따르는 확률변수로 해석한다. 즉 $(\alpha + u_i) \sim N(\alpha, \sigma_u^2)$
- 고정효과와 확률효과 중 어떤 모형을 선택할 것인지는 첫째로 u_i에 대한 해석에 달려있다
  - 패널 개체들이 모집단에서 무작위로 추출된 표본의 개념이라면 오차항 u_i는 확률분포를 따른다고 가정할 수 있다
  - 하지만 패널 개체들이 모집단에서 추출된 표본이 아니라 특정 모집단 그 자체라면 u_i는 확률분포를 따른다고 할 수 없다
  - 한국복지패널 자료를 분석할 때 모집단에서 추출하였기 때문에 확률효과를 사용하는 것이 바람직해 보인다
  - 이에 반해 OECD국가에 대한 분석이나 미국의 50개 주 패널에 대한 자료는 그 자체로 모집단을 형성하기 때문에 고정효과로 보는 것이 적절하다

하우스만 검정 2

- 통계적 측면에서 보았을 때, $cov(x_{it}, u_i) = 0$이면
  - 고정효과 추정량과 확률효과 추정량이 모두 일치추정량이다. 따라서 두 모형의 추정량이 유사할 것이다
- $cov(x_{it}, u_i) \neq 0$이면
  - 고정효과 추정량은 여전히 일치추정량이지만 확률효과 추정량은 일치추정량이 아니다. 따라서 두 추정량 사이에 차이가 존재할 것이다
- 하우스만(Hausman) 검정은 이러한 특성을 이용한다.
  - 즉 $H_0: cov(x_{it}, u_i) = 0$ 하에 두 추정량은 유사하지만 대안가설 하에 두 추정량은 차이가 있다
  - $H = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)' [var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})]^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \sim \chi^2_{K-1}$를 이용한다
  - 하우스만 검정에 있어 $var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$ 행렬이 양정행렬(positive definite)이 되어야 하지만 그렇지 않은 경우에도 큰 문제는 없는 것으로 알려져 있다

하우스만 검정 3

- phtest(lhwage ~ factor(union) +factor(disa) + factor(year) +
- age + agesq + factor(gend) + factor(marst) +
- jdur+factor(reg),
- data=padt, model=c("within","random"))

```
        Hausman Test

data:  lhwage ~ factor(union) + factor(disa) + factor(year) + age +  ...
chisq = 2243.3, df = 21, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

모형간 비교

OLS

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.826e+00  2.417e-02 365.189  < 2e-16 ***
factor(union)1  1.983e-01  6.062e-03  32.717  < 2e-16 ***
factor(union)2  2.772e-01  7.802e-03  35.532  < 2e-16 ***
factor(disa)1  -1.184e-01  1.030e-02 -11.493  < 2e-16 ***
factor(year)2013  6.714e-02  9.654e-03   6.955 3.56e-12 ***
factor(year)2014  7.707e-02  9.686e-03   7.957 1.80e-15 ***
factor(year)2015  1.444e-01  9.737e-03  14.827  < 2e-16 ***
factor(year)2016  1.917e-01  9.763e-03  19.632  < 2e-16 ***
```

Within

```
Coefficients: (1 dropped because of singularities)
                Estimate  Std. Error  t-value  Pr(>|t|)
factor(union)1  2.5597e-02  5.6620e-03   4.5209 6.176e-06 ***
factor(union)2  6.8102e-02  8.1773e-03   8.3282 < 2.2e-16 ***
factor(disa)1  -3.6206e-02  3.8193e-02  -0.9479   0.34316
factor(year)2013  9.8794e-02  6.7688e-03  14.5955 < 2.2e-16 ***
factor(year)2014  1.5795e-01  7.6475e-03  20.6540 < 2.2e-16 ***
factor(year)2015  2.4859e-01  8.8710e-03  28.0228 < 2.2e-16 ***
factor(year)2016  3.2617e-01  1.0269e-02  31.7640 < 2.2e-16 ***
factor(year)2017  4.3551e-01  1.1851e-02  36.7484 < 2.2e-16 ***
```

FD

```
Coefficients: (1 dropped because of singularities)
                Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)     9.7274e-03  4.3639e-03  2.2290  0.025817 *
factor(union)1  1.6255e-02  5.4533e-03  2.9808  0.002877 **
factor(union)2  4.4676e-02  8.2934e-03  5.3870 7.209e-08 ***
factor(disa)1  -2.9130e-02  5.0695e-02 -0.5746  0.565547
factor(year)2013  9.7230e-02  8.0483e-03 12.0808 < 2.2e-16 ***
factor(year)2014  1.5415e-01  1.3172e-02 11.7027 < 2.2e-16 ***
factor(year)2015  2.4004e-01  1.8070e-02 13.2836 < 2.2e-16 ***
factor(year)2016  3.2169e-01  2.2888e-02 14.0552 < 2.2e-16 ***
factor(year)2017  4.3449e-01  2.7752e-02 15.6561 < 2.2e-16 ***
factor(year)2018  5.2766e-01  3.2770e-02 16.1018 < 2.2e-16 ***
factor(year)2020  7.1218e-01  4.1391e-02 17.2063 < 2.2e-16 ***
factor(year)2021  7.9756e-01  4.6921e-02 16.9981 < 2.2e-16 ***
factor(year)2022  8.7388e-01  5.2927e-02 16.5112 < 2.2e-16 ***
agesq          -4.1719e-04  4.4146e-05 -9.4502 < 2.2e-16 ***
factor(marst)1  7.9579e-04  1.8993e-02  0.0419  0.966579
factor(marst)2 -9.8056e-02  3.5673e-02 -2.7487  0.005986 **
factor(marst)3  6.9168e-03  3.0337e-02  0.2280  0.819651
jdur            4.8592e-03  4.9524e-04  9.8118 < 2.2e-16 ***
factor(reg)1    2.4239e-02  2.3215e-02  1.0441  0.296458
factor(reg)2    6.8883e-03  1.9719e-02  0.3493  0.726853
factor(reg)3   -1.5641e-02  3.0748e-02 -0.5087  0.610980
factor(reg)4    2.1574e-02  4.7715e-02  0.4521  0.651172
```

Random

```
                Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)     8.6418e+00  3.0765e-02 280.8987 < 2.2e-16 ***
factor(union)1  7.9067e-02  5.3526e-03  14.7717 < 2.2e-16 ***
factor(union)2  1.5380e-01  7.6505e-03  20.1028 < 2.2e-16 ***
factor(disa)1  -1.1371e-01  1.6480e-02  -7.0213 2.198e-12 ***
factor(year)2013  6.7455e-02  6.5726e-03  10.2631 < 2.2e-16 ***
factor(year)2014  9.0938e-02  6.6737e-03  13.6264 < 2.2e-16 ***
factor(year)2015  1.5348e-01  6.7720e-03  22.6639 < 2.2e-16 ***
factor(year)2016  1.9956e-01  6.8492e-03  29.1370 < 2.2e-16 ***
factor(year)2017  2.7732e-01  6.9116e-03  40.1246 < 2.2e-16 ***
factor(year)2018  3.3915e-01  7.0068e-03  48.4032 < 2.2e-16 ***
factor(year)2020  4.5529e-01  7.2912e-03  62.4442 < 2.2e-16 ***
factor(year)2021  5.1109e-01  7.1339e-03  71.6423 < 2.2e-16 ***
factor(year)2022  5.5223e-01  7.2403e-03  76.2721 < 2.2e-16 ***
age             3.2803e-02  1.3516e-03  24.2691 < 2.2e-16 ***
agesq          -4.6048e-04  1.2982e-05 -35.4709 < 2.2e-16 ***
factor(gend)1  -2.9105e-01  8.0414e-03 -36.1945 < 2.2e-16 ***
factor(marst)1  1.4325e-01  9.7331e-03  14.7177 < 2.2e-16 ***
factor(marst)2  1.3715e-01  1.6987e-02   8.0737 6.819e-16 ***
factor(marst)3  2.7683e-02  1.5380e-02   1.8000   0.07187 .
jdur            1.5897e-02  3.6530e-04  43.5184 < 2.2e-16 ***
factor(reg)1   -7.1991e-02  1.0331e-02  -6.9685 3.203e-12 ***
factor(reg)2   -3.9134e-02  9.2711e-03  -4.2210 2.432e-05 ***
factor(reg)3   -9.7691e-02  1.3752e-02  -7.1040 1.212e-12 ***
factor(reg)4   -4.5359e-02  2.1025e-02  -2.1574   0.03097 *
```

Thank You

- 발표를 들어 주셔서 감사합니다
- 자료를 수집하고 배포하느라 고생하시는 한국복지패널 관계자님들께 감사합니다
- 질문이나 자료에 오류가 있다면 다음 이메일로 알려주세요
- sochyunsik@khu.ac.kr