2014 한국사회정책연합 공동학술대회 – 방법론 강의

패널 자료를 이용한 연구 방법

김 희 삼 (KDI 연구위원)

CONTENTS

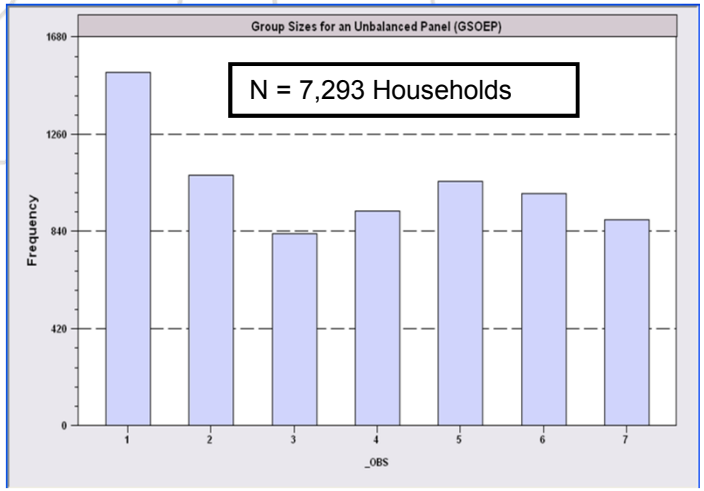
패널 자료의 의미

- **패널 자료(panel data):** 동일한 조사 대상 개체(예: 개인, 가구, 기업, 국가)들로부터 여러 시점에 걸쳐 반복적으로 수집한 자료
  - **패널 자료**는 동일한 개체에 대해 횡단면 자료와 시계열 자료가 통합(pooling)
    - **횡단면(cross-section)** 자료는 일정 시점에서 여러 개체들로부터 수집한 자료
    - **시계열(time series)** 자료는 시간의 흐름에 따라 수집한 자료
    - **합동 횡단면(repeated cross-section)** 자료는 동일한 개체를 반복적으로 관찰하는 것이 아니라 서로 다른 개체가 매 시점에서 조사된 시계열 자료(pseudo panel)
  - **패널 연구(panel study)**는 같은 주제에 대해서 시간 경과에 따른 변화를 연구하기 위해 반복적으로 관찰하는 **종단 연구(longitudinal study)**의 일종
    - 종단 연구에는 패널 연구, 추세 연구, 코호트 연구가 있음.
    - **추세 연구(trend study)**는 구성원은 변하지만 성격이 동일한 모집단(예: 현직 초등학교 교장)에서 상이한 표본을 상이한 시점에 조사하여 시점 간 관측치를 비교함으로써 모집단 내에서의 추세를 파악하는 연구
    - **코호트 연구(cohort study)**는 구성원이 변하지 않는 특정한 모집단(예: 2011년 마이스터고 입학생)으로부터 상이한 표본을 상이한 시점에서 표집하여 조사하는 연구

패널 자료의 형식

- 패널자료에서 개별 관측치는 조사 대상이 되는 **개체**(i, i=1,…,N)와 **관측 시점**(t, t=1,…,T)의 조합으로 이루어지므로 두 개의 하첨자(subscript)를 붙여서 표기
  - 예: x_{it}, y_{it}
- 복수의 시점에서 관측된 개체별 자료를 한 곳에 모아서 만든 패널자료는 **개체별 복수 시점** 관측치 또는 **시점별 복수 개체** 관측치를 가진 모양
  - 조사 대상 개체가 관측된 시점들이 모두 동일하면 **균형 패널(balanced panel)**이 되지만, 조사 누락 및 표본 탈락(attrition) 문제로 각 개체의 자료 포괄 기간이 달라져 **불균형 패널(unbalanced panel)**이 되는 경우가 일반적

가구ID(i)	조사연도(t)	가구총소득(x_{it})	가구총소비(y_{it})
1001	1998		
•	•		
•	•		
1001	2009		
1002	1998		
•	•		
•	•		
•	•		
1002	2010		



패널 자료 분석의 장점(1)
– 횡단면 분석과 시계열 분석의 한계 극복

연령(age) 효과와 코호트(cohort) 효과의 구분

- 예: <생활의 달인!> 입사연도와 경력연수에 따른 병아리 감별 생산성 차이

  - 생산성이 높은 고참 왈, "우리 고참들은 입사 때부터 달랐어! ~_~+"
  - 생산성이 낮은 신참 왈, "그냥 오래 하다 보니 잘 하게 된 거 아녜요? @.@;

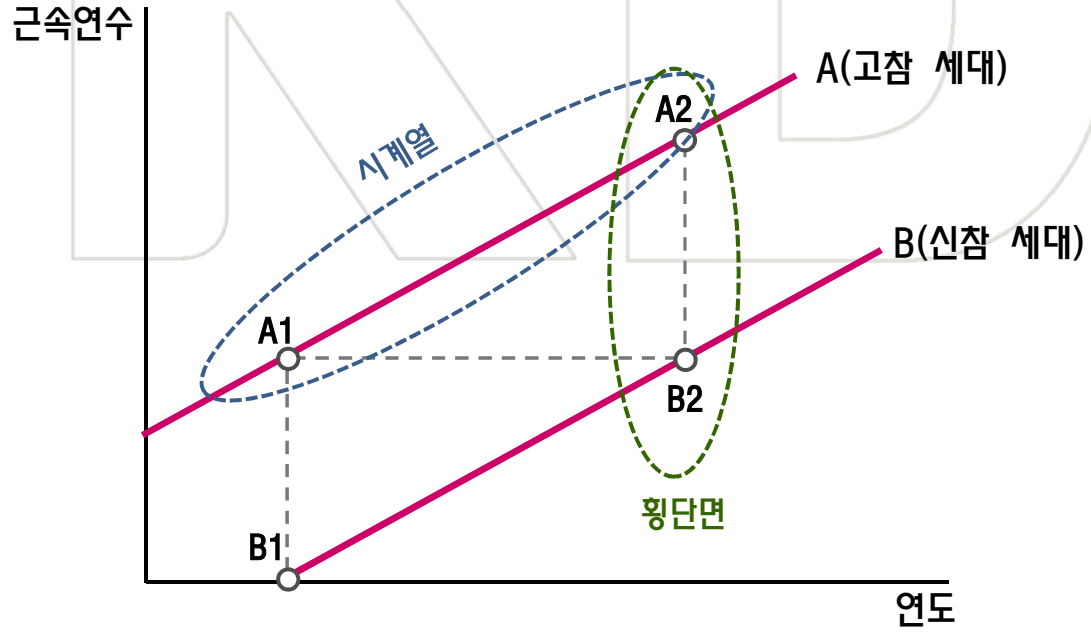
- 분석 모형: 시간당 병아리 감별 수 = α*입사연도 + β*근속연수 + 기타변인*γ + ε

- 횡단면 자료에서는 입사연도가 빠르고 근속연수가 긴 고참과 입사연도가 늦고 근속연수가 짧은 신참의 생산성 차이가 입사연도의 차이에 따른 코호트 효과인지 근속연수의 차이에 따른 연령 효과인지를 식별할 수 없음.
  - 특정 연도에 관측된 횡단면 자료에서 입사연도+근속연수=관측연도(일정)이므로 입사연도와 근속연수 간에 완전한 다중공선성(perfect multicollinearity) 존재!

패널 자료 분석의 장점(1)

– 횡단면 분석과 시계열 분석의 한계 극복

- 시계열 자료에서는 전후 연도 간 비교를 통해 연령(근속) 효과만 파악 가능
  - 고참 세대 시계열 자료에서 A1과 A2 시점의 생산성 비교, 또는 신참 세대 시계열 자료에서 B1과 B2 시점의 생산성 비교를 통해 근속에 따른 생산성 변화를 파악
- 패널 자료는 신참의 근속연수만큼 고참이 근속했을 당시의 고참과 현재의 신참을 비교함으로써 입사연도에 따른 생산성 차이, 즉 코호트 효과까지 파악할 수 있음
  - 고참의 A1 시점과 신참의 B2 시점의 생산성 비교를 통해 코호트 간 차이를 파악



패널 자료 분석의 장점(2)

– 이중차분비교를 통한 미관찰 이질성 통제

실험집단과 통제집단 간 이중비교를 통한 정책 효과 분석

- 예: 대학재정지원사업은 사업단 소속학과 졸업생의 취업률 제고에 효과가 있었나?

	사업단 1: 선정! ^^ 실험집단(처치집단) Treatment Group	사업단 2: 탈락! ㅠㅠ 통제집단(비교집단) Control Group
사업시행 전(Before)	취업률 T1	취업률 C1
사업시행 후(After)	취업률 T2	취업률 C2

- **단순비교를 통한 사업 효과 분석**
  - 횡단면 자료를 이용한 단순비교방법(Yardstick Method): T2-C2

    ※ 사업 효과뿐 아니라 사업단별 특성에 의한 차이를 포함할 수 있음.
  - 시계열 자료를 이용한 전후비교방법(Before & After Method): T2-T1

    ※ 사업 효과뿐 아니라 시기별 특성에 의한 차이를 포함할 수 있음.
  - 이러한 사업단별/시기별 특성은 자료에서는 관찰되지 않는 이질성(unobserved heterogeneity)로서 단순비교를 통해서는 통제(control)되지 않음.

패널 자료 분석의 장점(2)

– 이중차분비교를 통한 미관찰 이질성 통제

- **이중비교를 통한 사업 효과 분석: 이중차분(Difference-in-Differences: DD/DID)방법**
  - 방법 1: (T2-C2)-(T1-C1) = (사업 효과+사업단 특성)-(사업단 특성) = 사업 효과
    ※ 사업단별 특성이 있다면 이전 시기에도 있었을 것이라는 가정에 기반
  - 방법 2: (T2-T1)-(C2-C1) = (사업 효과+시기별 특성)-(시기별 특성) = 사업 효과
    ※ 시기별 특성이 있다면 두 사업단에 모두 있었을 것이라는 가정에 기반

- **이중차분방법에 상응하는 회귀분석모형:**

$$\text{취업률} = \alpha + \beta * \text{선정사업단 더미} + \gamma * \text{시행 후 시기 더미} + \delta * (\text{선정사업단 더미} * \text{시행 후 시기 더미}) + \varepsilon$$

  - 선정사업단 더미와 시행 후 시기 더미의 곱에 대한 추정계수 δ = 사업 효과
  - 회귀분석모형에서 기타 변인의 영향을 설명변수로 추가하여 통제할 경우 δ의 값은 달라질 수 있음.

패널 자료 분석의 장점(3)

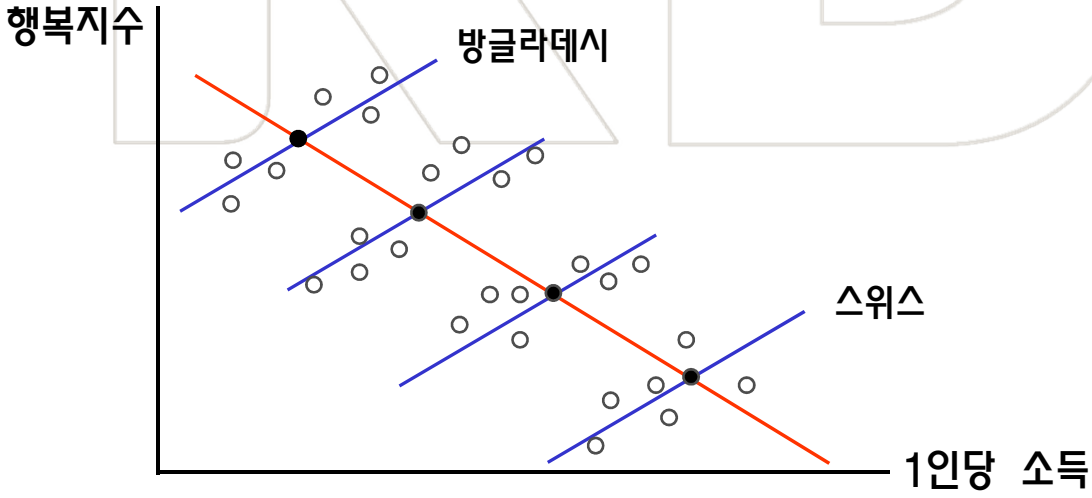
– 조사단위 간 평균 비교의 맹점 극복

조사 단위별로 고유한 내재적 속성(고정 효과)의 존재를 고려한 분석

- <행복의 역설?> 가난하면 더 행복한가? 쉽게 행복을 느끼는 나라는 가난해지나?

  1998년 런던정경대학(LSE)이 54개국을 대상으로 한 행복지수 조사결과, 행복도가 가장 높은 국가는 가난한 나라로 손꼽히는 방글라데시, 아제르바이잔, 나이지리아 순이었다. 미국이나 일본, 스위스 등은 40위권이었고, 한국은 23위로 이들 선진국보다 행복지수가 높았다.

- 같은 나라 내에서 부자와 빈자의 행복지수를 비교했다면?
- 같은 나라에 있어서 1인당 소득이 높은 연도와 낮은 연도의 행복지수를 비교했다면?



패널 자료 분석의 장점(4)
– 설명변수의 내생성 문제 완화

내생성(endogeneity) 문제가 있는 설명변수의 추정 편의(bias) 감소

- 예: 사교육의 진정한 효과는?
- (단순회귀)분석모형: 전교 석차 = α + β*사교육비 + ε
- 사교육비 지출에 영향을 주는 관찰되지 않은 개인별 특성(예: 능력)이 석차에도 영향을 준다면, 통상최소자승(OLS) 추정치에는 편의(bias)가 발생
  - 이 특성은 오차항(ε)에 반영되어 있을 것이므로 설명변수(X)인 사교육비와 오차항 간에 상관관계가 존재하게 되어(Cov(X,ε)≠0) OLS 추정량은 편의가 있고 일치 추정량도 되지 못함(biased and inconsistent)
  - 만약 능력이 높은(낮은) 학생이 사교육을 많이 받는다면 사교육의 효과는 과대추정(과소추정)
- 패널 자료에서 같은 학생이 해마다 사교육비 지출을 달리 했을 때 석차가 어떻게 변했는지를 분석하면, 관찰되지 않은 개인별 특성의 통제가 가능
  - 만약 개인별 능력이 해마다 달라지지 않고 일정한(time-invariant) 특성을 갖는 고정 효과(fixed effect)로 간주될 수 있다면, 이와 같은 방법으로 능력 변수의 누락에 따른 사교육 효과 추정의 편의를 제거할 수 있을 것임(but 희석편의 문제).

패널 자료 분석의 단점?

패널 자료 수집의 단점

- 일회성 조사나 다른 개체들에 대한 다시점 조사보다 같은 개체들에 대한 추적 조사는 더 큰 조사 비용과 관리 노력을 요구

패널 자료 분석의 단점

- 특정 개체를 반복적으로 조사할 경우 결측치가 발생할 가능성이 높아져 추정량의 효율성과 모수 식별(identification)에 문제가 생길 수 있음.
- 조사 기간이 길어지면 표본 탈락의 증가로 가중치 및 대표성 문제 발생
  - 그러나 고령자 패널 자료의 경우 사망 이탈자에 대한 출구 조사(exit survey)는 질병, 사망, 상속 등에 대한 유용한 정보 제공
- 대부분의 패널 자료(특히 개인 조사 자료)는 시간 변수의 길이(T)가 짧아 제한된 종속변수 모형(예: 패널 프로빗/로짓) 등의 추정 시 계산비용 발생

또 다른 단점?

- 쓸만한 자료가 없어 좋은 연구를 못 한다는 핑계를 더 이상 대기 어렵다!

기본적인 선형패널모형의 선택(1)

기본적인 선형패널모형

$$y_{it} = X_{it}^{'}\beta + u_i + \theta_t + \varepsilon_{it} \quad (1)$$

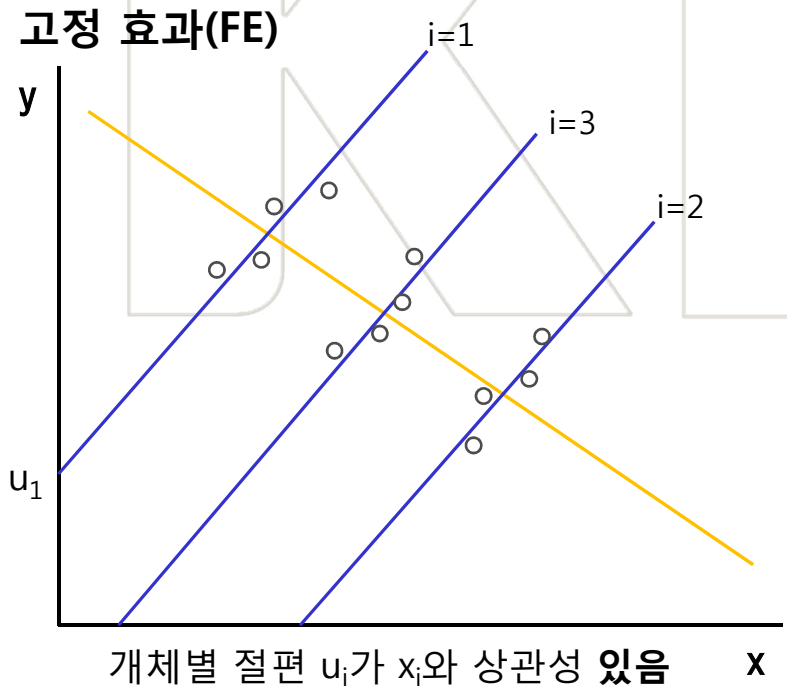
$$y_{it} = X_{it}^{'}\beta + u_i + \varepsilon_{it} \quad (2)$$

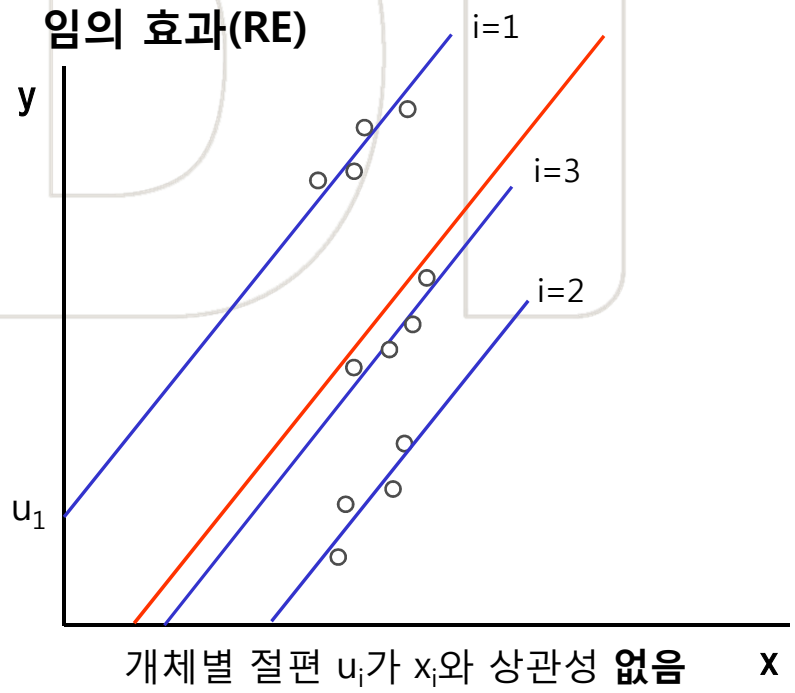
- 현실적으로 대부분의 패널 자료는 기간이 짧고 개체 수가 많은 형태
  - θ_t의 개수가 적은 상황에서는 시간 더미 변수를 넣는 것만으로 시간별 특수성은 고려할 수 있으므로, 모형 (1)보다는 모형 (2)에서 **관찰되지 않는 개체별 특수성 u_i를 어떻게 고려할 것인가**의 문제가 관건
- 관찰되지 않는 내재적 속성(u)와 관찰되는 속성(X) 간에 관련이 있는지에 따라 모형 선택이 달라짐.
  - u가 X와 관련이 없는 내재적 속성이면 **임의 효과(Random Effect: RE)** 모형
  - u가 X와 관련이 있는 내재적 속성이면 **고정 효과(Fixed Effect: FE)** 모형
- 대개는 조사대상의 관찰되지 않은 내재적 속성이 관찰된 속성과 관련이 있는 경우가 많아 고정 효과 모형을 사용하게 되는 경우가 일반적임.
  - 고정 효과가 존재하는 경우에 OLS 모형을 사용하면 개체의 고유한 속성으로 인한 공통성이 인과관계로 오인(spurious regression)될 수 있음(예: 행복의 역설).

기본적인 선형패널모형의 선택(2)

Pooled OLS, RE vs. FE의 선택

- **Pooled OLS를 써도 되는 경우:** LM 검정(Breusch & Pagan) 결과 u_i**의 분산이 0**일 때
  - X가 개체들의 특징을 모두 관측해서 설명해낼 수 있는 경우로서 u를 제거해도 무방
- **RE와 FE 중 FE모형을 써야 하는 경우:** Hausman 검정 결과 $Cov(X, u)=0$**이 기각**될 때
  - **FE**는 u를 각 개체 고유의 **상수항**으로 간주 (☞ 절편은 시간불변이나 개체별로 다르고, 기울기는 시간불변이며 개체간에도 동일)
  - **RE**는 u를 개체별로 우연히 주어진 **오차항**의 일부로 간주 (☞ 개체별 오차항의 다른 시점간 상관으로 인해 최소제곱추정 OLS 대신 일반화된 최소제곱추정 GLS를 사용)





[참고] 비실험적 분석방법의 선택편의 통제원리

분 류		선택편의(Selection Bias) 통제원리	해당 방법론
구조적 접근법 (Structural Approach)		프로그램 참여과정을 직접 모형화하여 관찰할 수 없는 특성들의 영향을 직접적으로 통제	**- Heckman의 2단계최소자승(2SLS) 모형**
관찰적 접근법 (Observational Study)		동일시점의 서로 다른 집단의 비교 또는 동일집단의 서로 다른 시점의 비교를 활용하여 프로그램의 효과를 추정	**- 횡단면비교방법(Yardstick Method)** **- 전후비교방법(Before & After Method)**
준실험적 접근법 (Quasi-experimental Approach)	**IV류 방법론**	프로그램 참여와 관련성이 높으면서 외생적 변량을 갖는 외부적 변수를 발굴하여 이 변수의 외생적 변화를 프로그램 효과추정에 활용	**- 도구변수(Instrumental Variable)모형** **- 회귀단절(Regression Discontinuity)모형**
	다시점 방법론	프로그램 참여를 결정짓는 내생적 요인들이 시간에 따라 일정하다고 가정하고 다른 시점에 측정된 데이터의 차분을 통해 관찰할 수 없는 요인들의 영향을 제거	**- 이중차분(Difference in Differences)모형** **- 고정효과(Fixed Effect)모형** **- 무작위성장(Random Growth)모형**
	매칭 방법론	선택편의가 관찰할 수 있는 특성들에 의해서만 발생한다고 가정하고, 특성변수의 범위와 분포의 차이를 통계적 방법에 의해 최대한 제거	**– 성향점수매칭(Propensity Score Matching)** **– 성향점수가중최소자승(Propensity Score Weighted Least Squares)모형**

[참고] 패널 자료의 시계열 성격을 활용한 분석

생존 분석(survival analysis, hazard model)

- 특정 상태의 지속기간과 외생적인 정책변수나 통제변수들 간의 관계 분석
  - 예: 국민기초생활보장제도 수급자의 탈수급/탈빈곤에 소요되는 시간, 실업급여 수급자의 특성(급여액이나 인적 특성)에 따른 탈실업/(재)취업에 소요되는 시간 등
- 오차항 정규분포를 가정하는 단순회귀분석은 지속기간의 분석에는 부적합
  - (1) 만약 탈출확률이 기간 내 일정하다면 지수분포, (2) 실업급여 수령자의 경우 실업 탈출이 수령초기나 수령말기에 집중된다면 지속기간의 최빈점은 1개가 아니라 2개, (3) 지속기간은 (-)가 될 수 없음, (4) 지속기간의 우측절단(right censoring: 조사시점에서 탈출하지 않은 사람의 지속기간은 중간에 잘린 채 관찰된 것) 등

(그래인저) 인과성 분석(Granger causality test)

- 상관성이 있는 두 변수 X와 Y 중에서 어느 것이 각각 원인이고 결과인가?
  - 예: 과거에 팔레스타인과 이스라엘 간의 잦은 무력분쟁은 어느 쪽의 공격이 원인인가? (-) 상관성을 보이는 사업장 내 내국인력과 외국인력의 고용은 내국인력의 부족을 외국인력이 메운 것인가, 내국인력을 내보내고 외국인력을 채용한 것인가?

$$Y_t = \beta_0 + \sum_{j=1}^{J} \beta_j Y_{t-j} + \sum_{k=1}^{K} \gamma_k X_{t-k} + u_t \,, \quad X_t = \beta_0 + \sum_{j=1}^{J} \beta_j X_{t-j} + \sum_{k=1}^{K} \gamma_k Y_{t-k} + u_t$$

패널 자료의 모습

. use http://www.stata-press.com/data/r10/nlswork

```
. list idcode year age race grade ln_wage hours, sepby(idcode )
```

	idcode	year	age	race	grade	ln_wage	hours
1.	1	68	.	.	.	.	.
2.	1	69	.	.	.	.	.
3.	1	70	18	2	12	1.451214	20
4.	1	71	19	2	12	1.02862	44
5.	1	72	20	2	12	1.589977	40
6.	1	73	21	2	12	1.780273	40
7.	1	75	23	2	12	1.777012	10
8.	1	77	25	2	12	1.778681	32
9.	1	78	26	2	12	2.493976	52
10.	1	80	28	2	12	2.551715	45
11.	1	82	.	.	.	.	.
12.	1	83	31	2	12	2.420261	49
13.	1	85	33	2	12	2.614172	42
14.	1	87	35	2	12	2.536374	45
15.	1	88	37	2	12	2.462927	48
16.	2	68	.	.	.	.	.
17.	2	69	.	.	.	.	.
18.	2	70	.	.	.	.	.
19.	2	71	19	2	12	1.360348	40
20.	2	72	20	2	12	1.206198	40
21.	2	73	21	2	12	1.549883	40
22.	2	75	23	2	12	1.832581	40
23.	2	77	25	2	12	1.726721	40
24.	2	78	26	2	12	1.68991	40
25.	2	80	28	2	12	1.726964	40
26.	2	82	30	2	12	1.808289	38
27.	2	83	31	2	12	1.863417	38
28.	2	85	33	2	12	1.789367	38
29.	2	87	35	2	12	1.84653	40
30.	2	88	37	2	12	1.856449	40
31.	3	68	22	2	12	1.493561	40
32.	3	69	23	2	12	1.702528	40
33.	3	70	24	2	12	1.451214	40

패널 자료의 구조 요약

```
. xtdes

  idcode:  1, 2, ..., 5159                                n =      4711
    year:  68, 69, ..., 88                                T =        15
           Delta(year) = 1 unit
           Span(year)  = 21 periods
           (idcode*year uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%      95%      max
                         1       1        3         5         9       13       15

     Freq.  Percent     Cum. |  Pattern
------------------------------+---------------------
       136     2.89     2.89 |  1.....................
       114     2.42     5.31 |  .....................1
        89     1.89     7.20 |  ..................1.11
        87     1.85     9.04 |  ....................11
        86     1.83    10.87 |  111111.1.11.1.11.1.11
        61     1.29    12.16 |  ..............11.1.11
        56     1.19    13.35 |  11...................
        54     1.15    14.50 |  ..............1.1.11
        54     1.15    15.64 |  .......1.11.1.11.1.11
      3974    84.36   100.00 |  (other patterns)
------------------------------+---------------------
      4711   100.00          |  XXXXXX.X.XX.X.XX.X.XX
```

패널자료를 요약해주는 명령어로는 xtsum 을 들 수 있다. 예컨대 노동시간을 살펴보자. 노동시간은 평균 36.56시간이다. 사람들간의 편차는 7.85시간이고, 사람 내에서의 편차는 7.52시간이다. 최대값과 최소값을 보고하고 있는데, 전체적으로 주당 1시간에서 168시간을 일하는 사람이 있다. 사람별 평균을 비교하면 적게는 1시간 많게는 83.5시간 일하고 있다. 사람 내에서는 평균으로부터의 deviation이 -2.15 시간부터 130.06시간임을 지칭한다.

```
. xtsum hours

Variable                |      Mean   Std. Dev.       Min        Max |   Observations
------------------------+--------------------------------------------+--------------
hours           overall |   36.55956   9.869623         1        168 |    N =   28467
                between |              7.846585         1       83.5 |    n =    4710
                within  |              7.520712  -2.154726   130.0596 |  T-bar = 6.04395
```

교육투자 수익률의 Pooling Estimates

```
. reg ln_wage grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure c.tenure#c.tenure 2.ra
> ce not_smsa south
```

Source	SS	df	MS
Model	2402.22796	10	240.222796
Residual	4011.63592	28080	.142864527
Total	6413.86388	28090	.228332641

Number of obs	= 28091
F(10, 28080)	= 1681.47
Prob > F	= 0.0000
R-squared	= 0.3745
Adj R-squared	= 0.3743
Root MSE	= .37797

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.0629238	.0010313	61.01	0.000	.0609024	.0649452
age	.038598	.003467	11.13	0.000	.0318025	.0453935
c.age#c.age	-.0007082	.0000563	-12.57	0.000	-.0008186	-.0005978
ttl_exp	.0211279	.002335	9.05	0.000	.0165511	.0257046
c.ttl_exp# c.ttl_exp	.0004473	.0001246	3.59	0.000	.0002031	.0006916
tenure	.0473687	.0019626	24.14	0.000	.0435219	.0512156
c.tenure# c.tenure	-.002027	.0001338	-15.15	0.000	-.0022893	-.0017648
2.race	-.0699386	.0053207	-13.14	0.000	-.0803673	-.0595098
not_smsa	-.1720455	.0051675	-33.29	0.000	-.182174	-.161917
south	-.1003387	.0048938	-20.50	0.000	-.1099308	-.0907467
_cons	.2472833	.0493319	5.01	0.000	.1505903	.3439762

교육투자 수익률의 Between Estimates

```
. xtreg ln_wage grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure c.tenure#c.tenure 2.
> race not_smsa south, be

Between regression (regression on group means)  Number of obs      =     28091
Group variable: idcode                          Number of groups   =      4697

R-sq:  within  = 0.1591                         Obs per group: min =         1
       between = 0.4900                                        avg =       6.0
       overall = 0.3695                                        max =        15

                                                F(10,4686)         =    450.23
sd(u_i + avg(e_i.))=  .3036114                  Prob > F           =    0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.0607602	.0020006	30.37	0.000	.0568382	.0646822
age	.0323158	.0087251	3.70	0.000	.0152105	.0494211
c.age#c.age	-.0005997	.0001429	-4.20	0.000	-.0008799	-.0003194
ttl_exp	.0138853	.0056749	2.45	0.014	.0027598	.0250108
c.ttl_exp# c.ttl_exp	.0007342	.0003267	2.25	0.025	.0000936	.0013747
tenure	.0698419	.0060729	11.50	0.000	.0579361	.0817476
c.tenure# c.tenure	-.0028756	.0004098	-7.02	0.000	-.0036789	-.0020722
2.race	-.0564167	.0105131	-5.37	0.000	-.0770272	-.0358061
not_smsa	-.1860406	.0112495	-16.54	0.000	-.2080949	-.1639862
south	-.0993378	.010136	-9.80	0.000	-.1192091	-.0794665
_cons	.3339113	.1210434	2.76	0.006	.0966093	.5712133

교육투자 수익률의 Random Effect Estimates

```
. xtreg ln_wage grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure c.tenure#c.tenure 2.
> race not_smsa south, re

Random-effects GLS regression                   Number of obs      =     28091
Group variable: idcode                          Number of groups   =      4697

R-sq:  within  = 0.1715                         Obs per group: min =         1
       between = 0.4784                                        avg =       6.0
       overall = 0.3708                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(10)      =   9244.74
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
```

ln_wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
grade	.0646499	.0017812	36.30	0.000	.0611589	.0681409
age	.0368059	.0031195	11.80	0.000	.0306918	.0429201
c.age#c.age	-.0007133	.00005	-14.27	0.000	-.0008113	-.0006153
ttl_exp	.0290208	.002422	11.98	0.000	.0242739	.0337678
c.ttl_exp# c.ttl_exp	.0003049	.0001162	2.62	0.009	.000077	.0005327
tenure	.0392519	.0017554	22.36	0.000	.0358113	.0426925
c.tenure# c.tenure	-.0020035	.0001193	-16.80	0.000	-.0022373	-.0017697
2.race	-.053053	.0099926	-5.31	0.000	-.0726381	-.0334679
not_smsa	-.1308252	.0071751	-18.23	0.000	-.1448881	-.1167622
south	-.0868922	.0073032	-11.90	0.000	-.1012062	-.0725781
_cons	.2387207	.049469	4.83	0.000	.1417633	.3356781
sigma_u	.25790526					
sigma_e	.29068923					
rho	.44045273	(fraction of variance due to u_i)				

교육투자 수익률의 Fixed Effect 추정치는 어디에?

```
. xtreg ln_wage grade age c.age#c.age ttl_exp c.ttl_exp#c.ttl_exp tenure c.tenure#c.tenure 2.
> race not_smsa south, fe
note: grade omitted because of collinearity
note: 2.race omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =     28091
Group variable: idcode                          Number of groups   =      4697

R-sq:  within  = 0.1727                         Obs per group: min =         1
       between = 0.3505                                        avg =       6.0
       overall = 0.2625                                        max =        15

                                                F(8,23386)         =    610.12
corr(u_i, Xb)  = 0.1936                         Prob > F           =    0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	(omitted)					
age	.0359987	.0033864	10.63	0.000	.0293611	.0426362
c.age#c.age	-.000723	.0000533	-13.58	0.000	-.0008274	-.0006186
ttl_exp	.0334668	.0029653	11.29	0.000	.0276545	.039279
c.ttl_exp# c.ttl_exp	.0002163	.0001277	1.69	0.090	-.0000341	.0004666
tenure	.0357539	.0018487	19.34	0.000	.0321303	.0393775
c.tenure# c.tenure	-.0019701	.000125	-15.76	0.000	-.0022151	-.0017251
2.race	(omitted)					
not_smsa	-.0890108	.0095316	-9.34	0.000	-.1076933	-.0703282
south	-.0606309	.0109319	-5.55	0.000	-.0820582	-.0392036
_cons	1.03732	.0485546	21.36	0.000	.9421496	1.13249
sigma_u	.35562203					
sigma_e	.29068923					
rho	.59946283	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(4696, 23386) =      5.13          Prob > F = 0.0000
```

Card, David, “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review, 43, 1990, pp. 245-257.

– 보트 난민의 유입이 지역 노동시장에 미친 효과(1)

1980년 5-9월 Miami에서 일어난 자연실험(natural experiment)

- Miami가 쿠바 난민이 밀려든 사건을 겪기 전후와 비교도시인 LA의 사건 전후 비교
- Difference in Differences 출동!

i = individual, T = 0 for no immigration, T=1 for migration

$(Y_i \mid T) = Y_{i,T} = 1$ if unemployed, 0 if employed.

c = city, t = period.

Unemployment rate in city c at time t is $E[Y_{i,0} \mid c,t]$ with no migration

Unemployment rate in city c at time t is $E[Y_{i,1} \mid c,t]$ with migration

Assume $E[Y_{i,0} \mid c,t] = \beta_t + \gamma_c$

$$E[Y_{i,1} \mid c,t] = \beta_t + \gamma_c + \delta$$

$$= E[Y_{i,0} \mid c,t] + \delta$$

$\delta =$ the effect of the immigration on the unemployment rate.

Card, David, “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review, 43, 1990, pp. 245-257.

– 보트 난민의 유입이 지역 노동시장에 미친 효과(2)

Applying the Model

- c = M for Miami, L for Los Angeles
- Immigration occurs in Miami, not Los Angeles
- T = 1979, 1981 (pre- and post-)
- Sample moment equations: $E[Y_i|c,t,T]$
  - $E[Y_i|M,79] = \beta_{79} + \gamma_M$
  - $E[Y_i|M,81] = \beta_{81} + \gamma_M + \delta$
  - $E[Y_i|L,79] = \beta_{79} + \gamma_L$
  - $E[Y_i|M,79] = \beta_{81} + \gamma_L$
- It is assumed that unemployment growth in the two cities would be the same if there were no immigration.

Card, David, “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review, 43, 1990, pp. 245-257.

– 보트 난민의 유입이 지역 노동시장에 미친 효과(3)

Implications for Differences

- If neither city exposed to migration
  - $E[Y_{i,0}|M,81] - E[Y_{i,0}|M,79] = \beta_{81} - \beta_{79}$ (Miami)
  - $E[Y_{i,0}|L,81] - E[Y_{i,0}|L,79] = \beta_{81} - \beta_{79}$ (LA)

- If both cities exposed to migration
  - $E[Y_{i,1}|M,81] - E[Y_{i,1}|M,79] = \beta_{81} - \beta_{79} + \delta$ (Miami)
  - $E[Y_{i,1}|L,81] - E[Y_{i,1}|L,79] = \beta_{81} - \beta_{79} + \delta$ (LA)

- In fact, one city (Miami) exposed to migration:
- The **difference in differences** is
  - $\{E[Y_{i,1}|M,81] - E[Y_{i,1}|M,79]\} - \{E[Y_{i,0}|L,81] - E[Y_{i,0}|L,79]\}$
    $= \delta$ (Miami)

Card, David, “The Impact of the Mariel Boatlift on the Miami Labor Market,” Industrial and Labor Relations Review, 43, 1990, pp. 245-257.

– 보트 난민의 유입이 지역 노동시장에 미친 효과(4)

DID Results

Table 6. Comparison of Wages, Unemployment Rates, and Employment Rates for Blacks in Miami and Comparison Cities.
(Standard Errors in Parentheses)

	All Backs				*Low-Education Blacks*			
	Difference in Log Wages, Miami − Comparison		*Difference in Emp./Unemp., Miami − Comparison*		*Difference in Log Wages, Miami − Comparison*		*Difference in Emp./Unemp., Miami − Comparison*	
Year	*Actual*	*Adjusted*	*Emp. − Pop. Rate*	*Unemp. Rate*	*Actual*	*Adjusted*	*Emp. − Pop. Rate*	*Unemp. Rate*
1979	−.15	−.12	.00	−2.0	−.13	−.15	.03	−.8
	(.03)	(.03)	(.03)	(1.9)	(.05)	(.05)	(.04)	(3.8)
1980	−.16	−.12	.05	−7.1	−.07	−.07	.03	−8.2
	(.03)	(.03)	(.03)	(1.6)	(.05)	(.05)	(.04)	(3.5)
1981	−.11	−.10	.02	−3.0	−.05	−.11	.04	−7.7
	(.03)	(.03)	(.03)	(2.0)	(.05)	(.05)	(.04)	(4.2)
1982	−.24	−.20	−.06	3.3	−.17	−.20	−.04	.6
	(.03)	(.03)	(.03)	(2.4)	(.05)	(.05)	(.04)	(4.7)
1983	−.21	−.15	−.02	.1	−.13	−.11	.04	−3.3
	(.03)	(.03)	(.03)	(2.7)	(.06)	(.05)	(.04)	(4.7)
1984	−.10	−.05	−.04	2.1	−.04	−.03	.05	.1
	(.03)	(.03)	(.03)	(2.4)	(.06)	(.05)	(.04)	(4.7)
1985	−.05	−.01	−.06	−5.5	.18	.09	.00	−4.7
	(.04)	(.04)	(.04)	(2.6)	(.07)	(.07)	(.06)	(5.6)

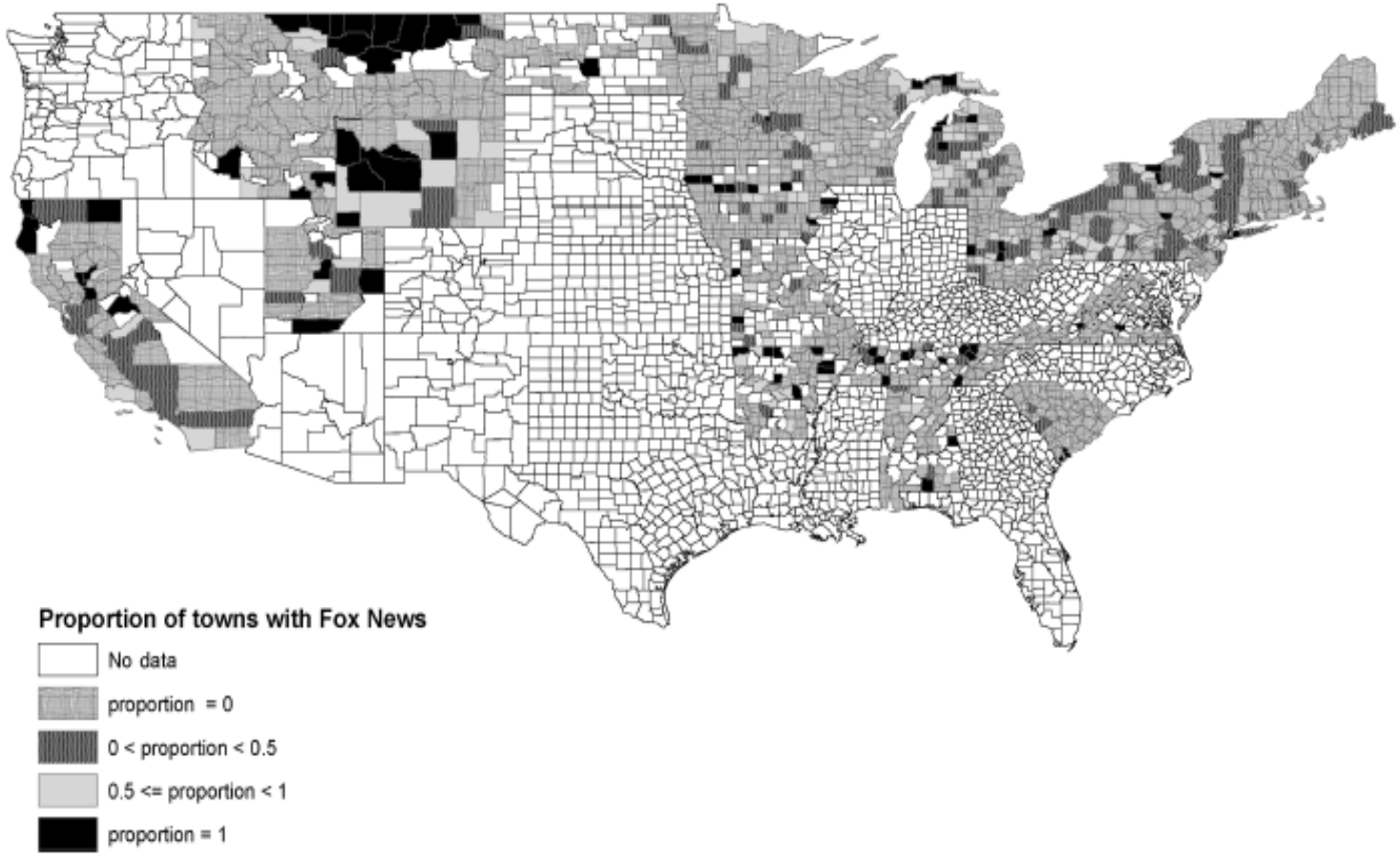
Notes: Low-education blacks are those with less than 12 years of completed education. Adjusted differences in log wages between blacks in Miami and comparison cities are obtained from a linear regression model that includes education, potential experience, and other control variables; see text. Wages are deflated by the Consumer Price Index (1980=100). “Emp.-Pop. Rate” refers to the employment:population ratio. “Unemp. Rate” refers to the unemployment rate among those in the labor force.

DellaVigna, Stefano and Ethan Kaplan,, “The Fox News Effect: Media Bias and Voting,” Quarterly Journal of Economics, 22 (3), pp. 1187-1234.

– 보수 성향 뉴스가 지역 유권자의 투표에 미친 효과(1)

1996년 10월~2000년 11월 중 일부 지역에 들어온 Fox News Channel

- 케이블 편성 채널에 보수(극우) 성향의 Fox 뉴스 채널 진입 현황은 지역별로 차이
  - 2000년 당시 county별 Fox 뉴스 채널 진입 현황



DellaVigna, Stefano and Ethan Kaplan,, “The Fox News Effect: Media Bias and Voting,” Quarterly Journal of Economics, 22 (3), pp. 1187-1234.

– 보수 성향 뉴스가 지역 유권자의 투표에 미친 효과(2)

- **9,256개 town의 투표 결과 자료를 이용하여 Fox 뉴스의 효과 분석**
  - 2000년까지 Fox 뉴스가 들어온 town들에서 공화당 득표율이 상승했는지 분석
- **Fox 뉴스 도입 전후인 1996년과 2000년 선거 결과를 DID 분석한 결과**
  - 대통령 선거에서 공화당은 Fox News가 들어온 지역에서 0.4~0.7%p 득표율 상승
  - 상원의원 선거에서도 공화당은 Fox News가 들어온 지역에서 시청자의 3~28%(시청률 측정도구에 따라 차이)를 공화당에 투표하도록 유도

– 특정 지역(집단)에 적용된 정책이나 환경 변화를 이용

최저임금 인상이 저임금노동자의 고용에 미친 영향

- 1992년 4월 뉴저지 주의 최저임금 인상($4.25→$5.05)이 패스트푸드점 고용에 미친 효과를 1992년 2월과 11월에 뉴저지 주와 인근(델라웨어강 건너) 펜실베니아 주 동부에서 수집한 고용 자료를 비교하여 분석(Card and Krueger, 1994, AER)

고용보호 판결이 비정규직 사용에 미친 영향

- 미국에서 고용주 임의고용(employment-at-will) 원칙의 예외를 인정하는 판결이 내려진 주에서 파견노동을 더 많이 사용하게 됐는지 분석(Auto, 2003, JoLE)

정보통신기술이 시장경제(일물일가법칙)의 작동에 미친 영향

- 인도에서 휴대폰 서비스의 도입이 지역 수산물시장들의 가격 편차를 줄였는지를 분석(Jensen, 2007, QJE)

종교활동 참여의 기회비용이 종교집회 출석에 미친 영향

- 미국에서 일요일 소매업 행위를 금지했던 법(Blue Law)이 폐지되는 시기의 주별 차이를 이용하여 예배 참여율에 미친 효과를 분석(Gruber and Hungerman, 2008, QJE)

Freeman, Richard, “Longitudinal Analyses of the Effect of Trade Unions,” Journal of Labor Economics, 3, pp. 1-26.

– 노조 가입이 임금에 미치는 효과

자료	횡단면(Pooled OLS) 추정치	고정효과(Fixed Effect) 추정치
5월 CPS, 1974-75	0.19	0.09
National Longitudinal Survey of Young Men, 1970-78	0.28	0.19
Michigan PSID, 1970-79	0.23	0.14
QES, 1973-77	0.14	0.16

‘횡단면 추정치 > 고정효과 추정치’의 결과 해석과 유의점

- 횡단면 추정치에 양(+)의 선택 편의가 존재한다! (관찰되지 않은 개인별 이질성 – 노조에 가입된 노동자가 더 높은 능력을 가졌을 가능성 - 을 고정효과 모형에서 통제하면 노조 가입의 임금 상승 효과는 줄어든다.)
- 그런데 고정효과는 측정오차로 인한 희석편의(attenuation bias)에는 취약하다! (노조 가입 여부는 지속성을 갖는 경우가 많아 개인의 시점간 변동이 작고, 특정 연도의 노조 가입 여부가 잘못 기재/보고되는 경우 노조 가입 여부의 연도별 변화는 노조 가입 여부 자체보다 잡음이 많은 정보이다. - Angrist and Pischke)

표본 연령대별 패널 자료의 현황

<table>
<tr><th rowspan="2">유아기</th><th colspan="3">청소년기</th><th rowspan="2">대 학</th><th rowspan="2">청 년</th><th rowspan="2">중 년</th><th rowspan="2">노 년</th></tr>
<tr><th>초등학교</th><th>중학교</th><th>고등학교</th></tr>
<tr><td rowspan="2">영유아
종단조사

아동패널</td><td rowspan="2">아동
청소년
패널

청소년
패널

서울교육
종단연구</td><td rowspan="2">청소년
패널

교육고용
패널

한국교육
종단연구

서울교육
종단연구</td><td rowspan="2">청년패널

교육고용
패널

한국교육
종단연구

서울교육
종단연구</td><td>청년패널

대졸자
직업이동
경로조사</td><td>청년패널

대졸자
직업이동
경로조사</td><td>고령화
연구패널</td><td>국민노후
보장패널

고령화
연구패널

한림고령자
패널

베이비부머
패널</td></tr>
<tr><td colspan="4">한국노동패널(만 15세 이상)
복지패널
사업체패널
산재보험패널
여성가족패널
여성관리자패널
의료패널
인구패널
인적자본기업패널
장애인패널
재정패널</td></tr>
</table>

연구기관별 패널 자료의 현황(1)

연구기관	조사명	시작 연도	조사 주기	표본 수	조사대상	표집방법	조사내용
국민연금연구원	국민노후보장 패널조사 (KReIS)	2005	격년	5,110가구	표본가구의 만50세 이상의 개인 및 그의 배우자 8,686명	확률 비례 층화	가구원 인적사항, 주거, 가구지출, 소득, 직업력 및 퇴직, 고용 및 취업, 고용 및 이직, 직업력 및 은퇴, 은퇴 및 구직, 노후준비, 공적 연금, 개인연금, 건강보험, 소득 및 이전지출, 자산 및 부채, 상속 및 증여, 가족관계, 돌봄, 삶의 만족도 및 건강
근로복지공단	산재보험 패널조사	2013	매년	2,000명	산재근로자	층화 비례추출법	인적특성, 산재서비스, 재해발생사업장, 현재 경제활동 판별, 재해 이후 경제활동, 건강 및 삶의 질, 개인소득, 가구 일반사항
서울대학교 노화 고령사회연구소	베이비부머 패널조사	2010	격년	4,668명	1955~1963년 출생자	-	사회인구학적 특성, 가족관계, 직업경력, 은퇴, 재정상황, 은퇴준비, 건강, 생애사건, 사회적 관계, 삶의 가치
서울특별시 교육연구정보원	서울교육 종단연구 (SELS)	2010	매년	16,500명	초 4학년 중 1학년 고 1학년	다단계 층화 표출	가정생활, 학교생활, 방과후 학교 및 사교육 참여, 학습시간, 자기주도적 학습태도, 여가활동, 건강, 가정배경, 교육비, 교사 배경 특성, 교사 근무시간, 학교장의 학교운영, 교육과정 편성 및 운영 등
육아정책연구소	아동패널조사	2008	매년	2,078명	표본가구의 0세 아동	층화 다단계 표본 추출법	대상 아동의 인구학적 특성, 건강특성, 발달 영역별 특성, 기질, 기초습관, 양육특성, 아버지 특성, 어머니 특성, 경제적 특성, 사회보장 지원 여부, 대리양육 여부, 기관.시설 향후 이용계획 등
한국고용정보원	대졸자 직업이동 경로조사 (GOMS)	2006	매년	26,000명	전문대, 교육대학, 4년제 대학 졸업자	다단계 층화 추출법	교육과정, 일자리, 재학 중 취업, 훈련과정, 자격증, 전공, 소득, 사회보험, 이직, 일자리 선택 시 중요도, 공무원 시험준비, 혼인상태, 거주지, 부모학력 등

연구기관별 패널 자료의 현황(2)

연구기관	조사명	시작 연도	조사 주기	표본 수	조사대상	표집방법	조사내용
한국고용정보원	청년패널조사 (YP)	2001	매년	5,956명	표본가구 중 만 15-29세에 해당하는 청년층 가구원	산업 직업별 고용조사에서 이중추출	학교생활, 교육평가, 사교육, 해외연수, 여가생활, 일자리, 구직활동 관련, 직업관 및 진로, 직업교육훈련 및 자격증, 진로결정성 및 경제활동여성, 직장체험, 인적 사항
		2007	매년	10,206명			※ 새로운 표본으로 시작 교육력, 아르바이트 및 휴학, 직업력, 구직활동 관련, 진로, 사교육, 해외연수, 문화적 자본, 직업교육훈련 및 자격증, 정부청년실업대책, 인적 사항
한국교육개발원	한국교육 종단연구 (KELS)	2005	매년	6,908명	2005년 전국 150개 중학교 1학년 학생	다단계 층화 군집추출법	[학생 수준] 인구통계적 특성, 개인적 특성, 수업요인, 학교생활, 교우관계, 가정생활, 진로, 수입, 직업능력수준, 직무만족도 [학교 수준] 학교특성, 교육평가, 교수활동, 교사활동, 상급학교 진학률, 학생수준 지표의 평균
한국노동연구원	사업체 패널조사	2006	매년	4,275개 사업장	상용근로자 30인(일반), 20인(공공) 이상	층화 추출	사업장 특성, 고용현황 및 고용 관리, 보상 및 평가, 인적자원관리 및 작업조직, 인적자원개발, 기업복지 및 산업재해, 응답자 정보, 노사관계
	한국고령화 연구패널조사 (KLoSA)	2006	격년	10,000명	45세 이상 중고령자 (제주도 제외)	집락층화 표집	인구, 가족(자녀, 손자녀, 부모, 형제자매), 건강상태, ADL과 간병인, 의료보장과 시설이용, 인지력, 신체기능, 고용, 소득, 자산, 주관적 기대감, 삶의 만족도 등
	한국 노동패널조사 (KLIPS)	1998	매년	6,700가구	표본가구의 15세 이상 가구원	2단계 층화집락 계통추출	가구원 정보, 사적 이전, 주거, 사교육, 소득, 소비, 저축, 자산, 부채, 경제활동상태, 주된 일자리 특성(근로계약, 기업형태 및 규모, 근로시간, 임금, 사회보험 등), 자격증, 학력, 혼인, 출산, 건강, 생활만족도 등

연구기관별 패널 자료의 현황(3)

연구기관	조사명	시작 연도	조사 주기	표본 수	조사대상	표집방법	조사내용
한국보건사회 연구원	의료패널조사	2007	매년	8,000가구	표본가구의 가구원	층화 추출	사회경제적 특성, 자산규모 및 생활비 지축, 의약품 구매, 노인장기요양, 인구사회경제적 특성, 경제활동 및 소득, 건강수준, 의약품 복용 형태, 병원 이용 목적 및 비용, 의료비 재원, 임신 및 출산, 생활습관, 의료접근성, 민간의료보험 등
	인구패널조사	2007	매년	6,000 ~ 10,000명	-	확률 비례 층화	교육, 직업, 건강, 소득, 주거 등
	한국복지 패널조사	2006	매년	7,000가구 (일반가구, 저소득층 가구 각각 3,500가구)	표본가구의 가구원	층화 이중 추출	가구원 일반사항, 보육, 교육, 건강 및 의료, 주거, 사회보험, 공공부조, 사회복지서비스, 소득, 지출 및 저축, 자산 및 부채, 직업이력, 경제활동상태, 고용지원프로그램, 생활실태 등
한국여성정책 연구원	여성가족 패널조사	2007	격년	10,000명	만 19~64세 이하의 여성이 가구원으로 거주하고 있는 가구와 해당 여성 가구원	다단계 층화 계통추출	가구원 및 가족사항, 가구소득 및 소비, 주거상태, 자산 및 부채, 성장과정 및 학교생활, 남편 일자리, 형제와의 관계, 첫 직장 경험, 장애인 및 환자, 혼인 상태, 출산경험과 자녀, 결혼과 부부생활, 현재의 경제활동, 사회보험, 차별사항, 일 만족도, 교육 및 훈련, 모성보호 제도 등
	여성관리자 패널조사	2007	격년	2,361명	100인 이상 기업에 종사하고 있는 대리급 이상 여성관리자	층화 2단 단순임의추출	현 직장사항, 인사관리, 경력개발, 취업력, 학력, 자격증, 교육훈련, 삶에 대한 만족도, 혼인, 배우자, 자녀

연구기관별 패널 자료의 현황(4)

연구기관	조사명	시작 연도	조사 주기	표본 수	조사대상	표집방법	조사내용
한국장애인 고용촉진공단	장애인 고용패널조사	2008	매년	5,000명	만 15~75세의 '장애인복지법' 제2조에서 규정하고 있는 유형의 장애인	층화 계통 추출	인적 사항, 장애유형 및 등급, 장애 발생시기, 원인, 상태, 경제활동, 직무수행능력, 전공/진로, 고용서비스, 직업능력개발, 건강, 운동, 수면, 여가, 임신, 출산, 양육과 일, 가사부담과 일, 은퇴, 노후준비, 근로소득, 사적 이전소득, 가구구성, 소득, 지출, 자산, 주거형태 등
한국조세 연구원	재정패널	2008	매년	5,010가구	-	2단계 집락 추출법	인적 사항, 주택 및 자동차, 지출, 소득이전, 자산 및 부채, 연금 및 보험, 소득세 납부
한국직업능력 개발원	인적자본 기업패널	2005	격년	450개 기업과 소속 14,000명 근로자	근로자 수 100인 이상이면서 일반 기업 이상	산업별, 규모별, 기업형태별	경영일반, HR부서, 인적자원관리, 인적자원개발, 인력현황, 연구개발, 기본정보, 현 직장 기본정보, 회사의 경쟁력 수준, 직무분석, 작업 환경 등
	한국 교육고용 패널조사	2004	매년	6,000명 (중학교, 일반계 고교, 전문계 고교 각 2,000명)	2004년 중학교, 일반계 고교, 전문계 고교 3학년 재학생	층화 집락 추출법	소속대학, 대학 만족도, 전공 및 공부, 성적, 진학 계획, 졸업 후 진로계획, 구직활동, 직업경험, 현재의 직업생활, 취업의사 및 준비, 영어 교육, 제2외국어 교육, 공무원시험 준비, 교양/취미 교육, 가구형태와 구성원, 가구 소득 및 지출, 여가생활
한국청소년 정책연구원	청소년 패널조사	2003 (중2) 2004 (초4)	매년	3,697명(중2) 2,949명(초4)	청소년 및 부모	층화 다단계집락 표집	[학생] 인적 사항, 직업선택, 향후 진로설정, 진로준비, 여가, 생활영역별 시간배분 및 중요도, 자아관 등 [부모] 가족구성, 동거여부, 학력, 직업, 근로형태, 소득, 사교육비, 주거형태 등