2009년도 제4차
「저출산고령사회 포럼」

□ 일시 : 2009. 12. 29(화) 15:00~18:00

□ 장소 : 한국보건사회연구원 신관 제2회의실

2009년도 제4차「저출산고령사회 포럼」(안)

▣ 일　　시: 2009. 12. 29(화) 15:00~18:00

▣ 장　　소: 한국보건사회연구원 신관 제2회의실

▣ 프로그램:

좌장: 정경희(한국보건사회연구원 저출산고령사회연구실장)

15:00~16:30 **주제발표**

발표: "다수준기법(Multilevel Analysis Technique)의 기본 원리와 출산력분석에서의 응용"

조영태(서울대학교 보건대학원 교수)

16:30~16:40 **coffee break**

16:40~17:10 **지정토론**

신윤정(한국보건사회연구원 저출산고령사회연구실 부연구위원)
박종서(한국보건사회연구원 저출산고령사회연구실 선임연구원)

17:10~18:00 **종합토론 및 폐회**

다수준분석기법(Multilevel Analysis Technique)의 기본 원리와 출산력분석에서의 응용

다수준분석기법(Multilevel Analysis Technique)의 기본 원리와 출산력분석에서의 응용

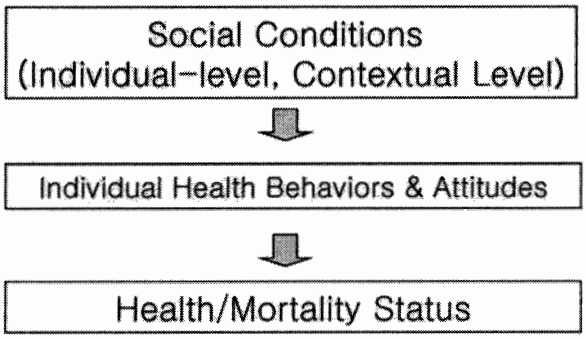
조 영 태
(서울대학교 보건대학원)

이론적 배경

- Society or community is not just an aggregate of individuals
- Individuals always interact with the social contexts to which they belong
  - Individuals are influenced by the social groups they belong
  - The properties of those groups are also influenced by individuals
- Sociology, in nature, deals with hierarchically structured study subjects
  - Eg.) Social forces, Cultural relativism, Structural Functionalism, etc.

이론적 배경- 보건사회학의 예

- 개인 건강/질병의 위험요소로서 환경의 중요성
- Social Conditions as Fundamental Causes of Disease (Link and Phelan 1995)



이론적 배경- 보건사회학의 예

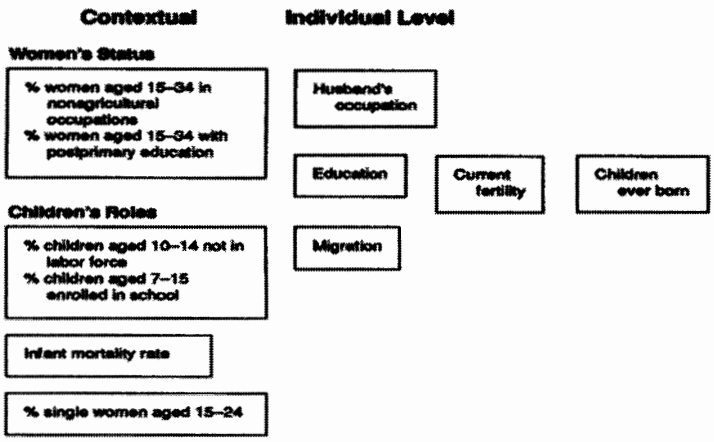
- 다양한 개인 건강/질병 outcome에 대한 환경(지역) 특성의 독립적인 효과가 보고됨 (Yen & Syme 1999).

- 지역 특성 (Robert 1999)
  - 사회경제적 환경
  - 서비스 환경
  - 물리적 환경

  } 집단적 가치관, 태도, 사회적 자본,

  ⇒ 개인의 건강/질병

이론적 배경 – 출산력 분석의 예



(Hirschman and Guest 1990 Demography)

한국 출산력분석에서의 가능성

- So far, mostly focused on individual level characteristics
- But it is certain that each ku/kun has different level of fertility, fertility intention, or fertility related values
- It may be attributable not only to the individual's compositions but also to the area's own contexts.

방법론적 배경

- Hierarchically Structured Data
  - Multi-stage sampling
    - Individuals, nested in Primary Sampling Units, nested in Strata.
    - 전국출산력조사: 시/도-구/군-동/읍,면
  - 같은 조사구를 가진 개인들은 어떤 변수들에 대해 서로 공유하는 공통적인 면모를 지닐 수 있는 반면 그것이 다른 조사구의 개인들과는 다를 수 있다.
    - Non-independence or autocorrelation
    - 기존의 회귀분석은 모든 개인이 무작위적인 지역 분포를 가진 것으로 가정한다.

왜 다수준 분석법인가?

(주어진 상황: individuals nested within areas with both individual-and area-level characteristics)

- 선택1: To ignore the macro-level units and attributes
  - Autocorrelation
  - Individualization
- 선택2: To aggregate individuals to the macro-units
  - No individuals
  - Ecological fallacy

왜 다수준 분석법인가?

- 선택3: Separate regressions for each area
  - Ignores the macro-level characteristics
  - Not practical (when the number of group is large)
- 선택4: Contextual analysis (지역변수와 개인변수를 같은 수준에 있다고 가정함)
  - Ignores the groupings
  - Autocorrelation
  - Assumes invariant effects of individual and group level characteristics

왜 다수준 분석법인가?

- 선택5: Analysis of Covariance Analysis (ANCOVA, 지역간 차이를 보기 위한 지역 dummy)
  - Ignores the group effects
  - Assumes equal individual-level effects across groups
  - Not practical (99 dummies when 100 areas)
- 선택6: Contextual + ANCOVA
  - Assumes equal individual and group level effects across groups
  - Not practical

왜 다수준 분석법인가?

- 선택7: Intercept (or slope) as outcome model (varying coefficient model)
  - 각 지역에 개인수준의 변수들로 각각의 regression 분석 (예, 100개의 regression equations).
  - 100개의 coefficient들을 종속변수로 놓고 지역변수들이 독립변수가 되어 2차 regression분석을 실시함.
  - 지역수준 변수의 효과를 보는데 유용함.
  - 개인변수의 지역별 효과를 볼 수 있지만 공통적인 효과를 볼 수 없음.
  - 어떤 지역에서는 개인변수의 효과가 통계적으로 무의미 할 수도 있지만 (large standard error), 2차 회귀분석 시 무시됨.
  - Non-practical.

다수준 분석법의 유용성

- 개인(micro)수준 변수와 지역(macro)수준 변수를 동시에 모델에 포함시킨다.
- 한 지역에 포함된 개인들이 서로 연관적일 가능성을 고려할 수 있다.
- 출산력의 지역간 차이를 간단한 parameter로 확인할 수 있다.
- 그 차이의 원인이 compositional한 특성에서 기인한 것인지 contextual한 특성에서 기인하는 것인지 확인 가능하다.
- 다수준 구조를 가진 대부분의 데이터에 적용 가능하다. (예, 학생-학교, 환자-병원, 종단자료 (Hazard model, SEM), APC model)

다수준 분석법의 특징

- 다양한 명칭
  - Multilevel Analysis, Hierarchical Linear Model (HLM), Mixed Model, Random Effect (Coefficient) Model, Variance Component Model.
- 교재마다 다른 notations
  - level 2 random variance: $\sigma^2 \sim \tau^2$
  - Level 1 error term: $R_{ij} \sim e_{ij}$
- Soft Wares: HLM, SAS, SPSS, MLWin, MIXREG
  - HLM http://www.ssicentral.com/
    - Full version, Rental license, student version

Basic Model (Two Level Case)

$$Y_{ij} = b_{0j} + b_{1j}X_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Note the j subscript, which denotes group identification

Y_{ij} = dependent variable for *i*th individual in *j*th group
X_{ij} = individual-level independent variable for *i*th individual in *j*th group.
ε_{ij} = individual-level error term for *i*th individual in *j*th group, normally distributed with mean of 0 and variance of σ^2.

- Unlike conventional modeling techniques, where coefficients (intercept and slopes) are assumed to be fixed, this model allows them to vary across groups.

$$b_{0j} = \gamma_{00} + \gamma_{01}C_j + U_{0j} \qquad U_{0j} \sim N(0, \tau_{00})$$

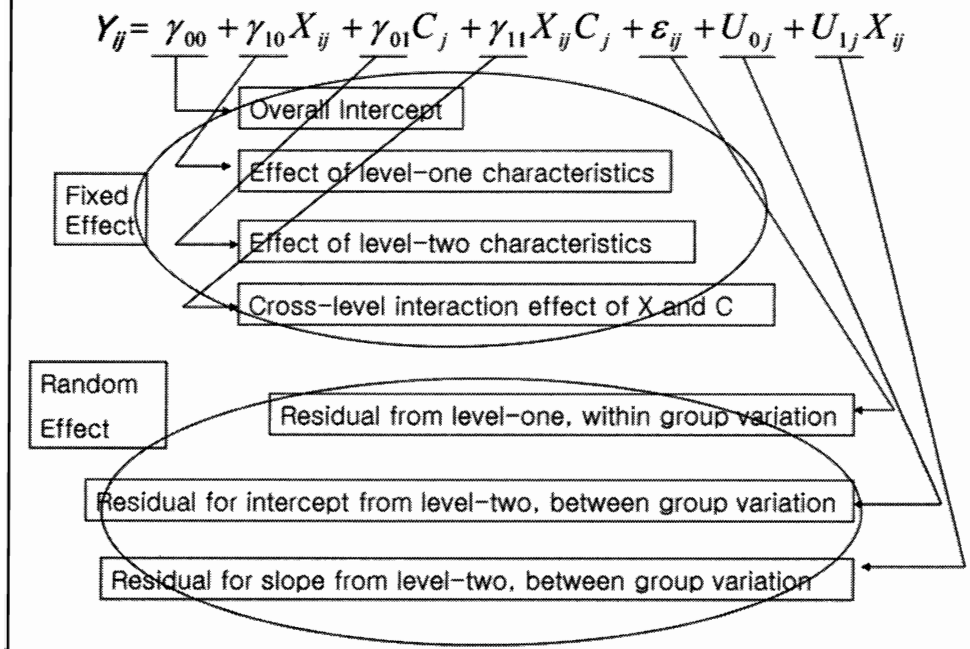
$$b_{1j} = \gamma_{10} + \gamma_{11}C_j + U_{1j} \qquad U_{1j} \sim N(0, \tau_{11})$$

$$\text{COV}\,(U_{0j},\ U_{1j}) = \tau_{10}$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}C_j + U_{0j} + (\gamma_{10} + \gamma_{11}C_j + U_{1j})X_{ij} + \varepsilon_{ij}$$

$$= \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}C_j + \gamma_{11}X_{ij}C_j + \varepsilon_{ij} + U_{0j} + U_{1j}X_{ij}$$

C_j = group-level independent variable
γ_{00} = overall intercept
γ_{01} = effect of group-level independent variable on intercept
γ_{10} = overall slope
γ_{11} = effect of group-level independent variable on slope
U_{0j} = deviation from overall intercept for each group
U_{1j} = deviation from overall slope for each group
τ_{00} = group-level variance of intercept
τ_{11} = group-level variance of slope
τ_{10} = covariance of intercept and slope
(if positive, when intercept increases, slope also increases)



- Thus, multilevel model is different from conventional regression models in that it includes both fixed and random coefficients.

- Also, individual- and group-level effects are simultaneously estimated.

- Error term is now decomposed to within- and between-group variances.

- Due to random parts, iterative algorithm should be used to generate population estimates.

- If τ_{00} or τ_{11} are statistically equal to zero, this model becomes the same as conventional model.

.... this means that intercept and slope are not random or do not vary across groups. "They are fixed"

- Therefore, decision for the use of multilevel analysis starts from analyzing the random variance for intercept from null model.

Modeling Strategy

1. Calculate the intra-class correlation coefficient (ρ), using intercept only model.

- Intra-class correlation coefficient
  - A measure of the degree of dependence of individuals.
  - It tells us the extent of error variance associated with groups.

  "the more individuals share common experiences due to closeness in space and/or time, the more they are similar, and the higher the intra-class correlation."

- The proportion of the variance in the outcome variable that is between the second- level units.

Modeling Strategy

- If there is significant intra-class correlation, it means individuals nested in a group exhibit significant autocorrelation, which makes the conventional methods not useful.

- then how to calculate ρ:

$$\rho = \frac{population\_variance\_between\_marco\_units}{total\_variance} = \frac{\tau}{\tau + \sigma}$$

- Here if τ is not statistically different from zero, ρ will be zero as well, which means no error variance is attributable to groups.

- In this case, we do not need to use multilevel analysis.

Model 1: $\rho = 0.58$, with significant random intercept variance.

$0.87/(0.64+0.87) = 0.58$

⇨ 58 percent of total variance is attributable to group-level variations.

Modeling Strategy

2. Progressively include individual-level independent variables, paying attention to random intercept variance (τ_{00}).

   - if the value of τ_{00} does not substantially change in its magnitude and significance, there exit group differences independent of individual characteristics.

   - this indicates the need of further investigation of contextual characteristics.

   - if the value of τ_{00} decreases or becomes not significant with inclusion of individual level characteristics, this means the area variation is attributable to the composition of individuals.

Modeling Strategy

3. Include group-level independent variables as well as individual-level variables.

   - if the value of τ_{00} decreases, group differences in dependent variable is attributable to the group-level variables.

   - if inclusion of group-level variables does not change the effect of independent variables, group-level variables and individual-level variables have independent effects.

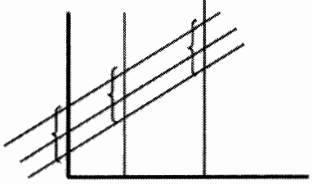
Modeling Strategy

4. If it is suspected that the effect of any individual level characteristics on dependent variable varies across groups, let the individual level variable to be random.

   - here, we look at the value and significance of τ_{11}.
   - when slope is allowed to be random, it is advised to utilize group or grand centering of the individual variable.
     # group mean centering: $(x_{ij}-x_{\cdot j})$
     # grand mean centering: $(x_{ij}-x_{\cdot\cdot})$

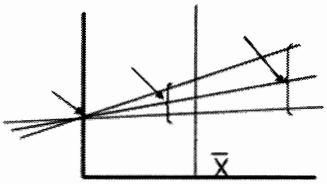
Centering

- 일반적으로 intercept는 모든 독립변수가 0일 때, 종속변수의 기대값 혹은 평균을 의미한다.
- 이는 다수준분석에서 intercept만 random하게 설정되어 있을 때도 마찬가지이다.
- 이 경우 intercept의 random variance는 독립변수가 어떤 값의 취해도 변하지 않는다.



Centering

- 하지만 만일 slope도 random하게 설정되면 독립변수가 어떤 값을 취하는가에 따라 intercept의 random variance의 값은 크게 변할 수 있다.
- 이 때 만일 독립변수가 0일 때 종속변수의 기대값이 실질적인 의미가 없는 경우, intercept의 random variance 역시 무의미한 정보를 전달하게 된다. (예, 소득과 건강... 소득이 0일 때 건강수준?)
- 그러므로 intercept를 실제 데이터에서 독립변수의 가장 대표적인 곳에 위치시키면, 그 때 intercept값과 그 random variance가 실질적인 정보를 전달하게 된다.



Modeling Strategy

4. Random slope 의 연장

   - you can have any independent to be random across groups (even dichotomous ones)

   - when random intercept and slope are both considered simultaneously, you need to analyze the covariance τ_{01}.

- 실제 분석을 수행하는데 있어 Random slope모델은 잘 사용되지 않는다. 그 이유는 이론적이기 보다는 기술적인 면이 강한데, slope을 random하게 설정하면 컴퓨터 패키지가 모수를 추정하는 시간이 길어지며 때로는 MLE의 iteration과정에서 convergence가 이루어지지 않기도 한다. 만일 convergence가 이루어져도 해석하는데 간단하지 않으므로, 많은 경우 random intercept model로 분석을 마치게 된다.

Modeling Strategy

5. You can include cross-level interaction terms in the model (individual*group).

   - cross-level interaction terms can make the random slope variance disappear.

6. For model fit, REML algorithm generates -2ResLL, which is analogous to -2LL (SAS).

Modeling Strategy

7. To be or Not to be Random
   - each predictor may be assigned to be random,
   - each random slope may covary with any other random slopes.
   - but parsimonious model is more desirable...
   - then what is a good guide for a fixed or a random slope?
   - in general, coefficients with strong fixed effect..
     the chance of varying slope is high..
   - but it is also possible.... a coefficient is not significant,
     and it is due to varying effect of the variable across macro-units.

- +++> Theory!!!

다수준 분석의 실례

1. 종속변수가 continuous 변수일 때.

Popularity Data (Hox 2002 Multilevel Analysis: Techniques and Applications)

Schools (N=200)

Teacher Experience (Z, in years)

Pupils (N=2000)

Y: Popularity (in a self-rating scale,
0 very unpopular -10 very popular)

X: Gender (0=boy, 1=girl)

Table. Multilevel Analysis Result. The effect of student gender and teacher experience on student's popularity

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	Coeffi.	S.E	Coeffi.	S.E	Coeffi.	S.E	Coeffi.	S.E	Coeffi.	S.E	Coeffi.	S.E
Fixed Effect												
Intercept	5.31	0.10	4.90	0.10	3.56	0.17	3.34	0.16	3.34	0.16	3.31	0.16
Level-1												
Gender[Boy]												
Girl			0.84	0.03	0.84	0.03	0.84	0.06				
Centered Gender									0.84	0.06	1.33	0.13
Level-2												
TeaExp					0.09	0.01	0.11	0.01	0.11	0.01	0.11	0.01
Cross-level Interaction												
Cgender*TeaExp											-0.03	0.01
Random Effect												
R (level-1)	0.64	0.02	0.46	0.01	0.46	0.01	0.39	0.01	0.39	0.01	0.39	0.01
U0 (Intercept)	0.88	0.13	0.86	0.13	0.49	0.07	0.41	0.06	0.41	0.06	0.41	0.06
U1 (Gender)							0.02	0.04	0.02	0.04	0.02	0.04
U11(Covariance)							0.27	0.05	0.27	0.51	0.23	0.04
-2ReLL	5115.60		4492.90		4444.40		4275.90		4275.90		4268.40	
Chi^2 Test (DF)			622.7(1)		48.5(1)		168.5(2)		N/A		7.5(1)	

SAS와 HLM의 차이점

- SAS는 level-one 데이터와 level-2 데이터가 하나의 data set으로 구성되어야 한다.

SAS Data Structure

Obs	PUPIL	SCHOOL	POPULAR	SEX	TEXP
1	1	1	8	1	24
2	2	1	7	0	24
3	3	1	7	1	24
4	4	1	9	1	24
5	5	1	8	1	24

- HLM은 하나의 데이터 혹은 각 level이 따로 구성된 데이터 모두를 사용할 수 있다.
  - 이 경우 다른 두 수준의 데이터를 연결해주는 변수를 "ID" 변수로 지정해야 한다.

SAS와 HLM의 차이점

- SAS는 종속변수가 continuous 변수일 때, proc mixed 프로시져를 사용하는데, model statement에 두 level이 합쳐진 equation을 포함시켜야 한다
- HLM은 각 level을 교과서적인 equation을 통해 연결시킨다 – SAS에 비해 간단함.
- Centering 도 SAS는 Centered된 변수를 data 단계에서 포함시켜야 하는 반면 HLM은 프로그램에 centering option이 있다.
- Model Fit: SAS는 -2 Res Log Likelihood를 HLM은 Deviance값을 산출한다. 둘의 차이는 MLE방법의 차이에서 비롯되는데 큰 차이가 없다.

출산력분석에서 다수준분석의 이론적 배경과 관련하여 도움이 될만한 논문들

- DiPrete, T.A., and J. Forristal. 1994. "Multilevel Models: Methods and Substance." *Annual Review of Sociology* 20: 331-357.
- Courgeau, D., and B. Baccaini. 1998. "Multilevel Analysis in the Social Science." *Population: An English Section* 10: 39-71.
- Robert, S. A. 1999. "Socioeconomic Position and Health: The Independent Contribution of Community Socioeconomic Context." *Annual Review of Sociology* 25: 489-516.
- Hirschman and Guest. 1990. "Multilevel models of fertility determination in four Southeast Asian Countries." Demography 27: 369-396.
- McNay et al. 2003. "Why are uneducated women in India using contraception? A multilevel analysis." Population Studies 57: 21-40.
- Example1, example 2

유용한 다수준분석 교재

Hox, Joop. 2002. Multilevel Analysis: Techniques and Applications

Snijders, Tom and Roel Bosker. 1999. Multilevel Analysis: An introduction to basic and advanced multilevel modeling

Kreft, Ita and Jan de Leeuw. 1998. Introducing Multilevel Modling

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. Hierarchical Linear Models: Applications and Data Analysis Methods.

MEMO

MEMO

MEMO

MEMO

MEMO