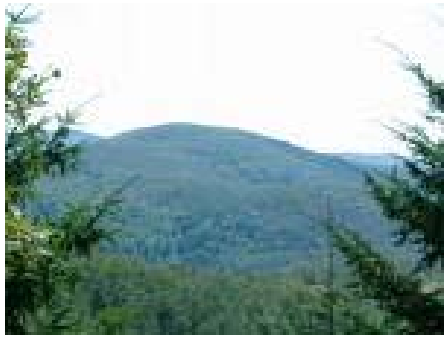
다층모형을 사용한 분석

서울대학교
언론정보학과
강남준

다층모형(Multi-level Modeling)



숲도 보고 나무를 봐라 !

- 위계적 선형모형(hierarchical linear model),
- 랜덤계수모형(random coefficient model)
- 기울기예측모형(slope-as-outcome model)

- 개인들의 관측 값이 시간에 따라 반복적으로 측정되었을 경우 (이를테면, 패널 자료), 개체성장곡선모형(individual growth curve model) 또는 개체변화모형(individual change model) 등으로 지칭되기도 한다

- 다양한 컴퓨터 패키지: HLM, MLwiN, AMOS, LISREL MPlus

다층모형과 일반회귀분석(OLS)의 비교

- 다층모형에서는 '개념적,' '이론적' 혹은 '경험적'으로 관측 값들의 측정 수준이 구별된다. 반면, 일반적 회귀분석에서는 모든 관측 값들은 하나의 측정수준을 갖는다고 가정한다.

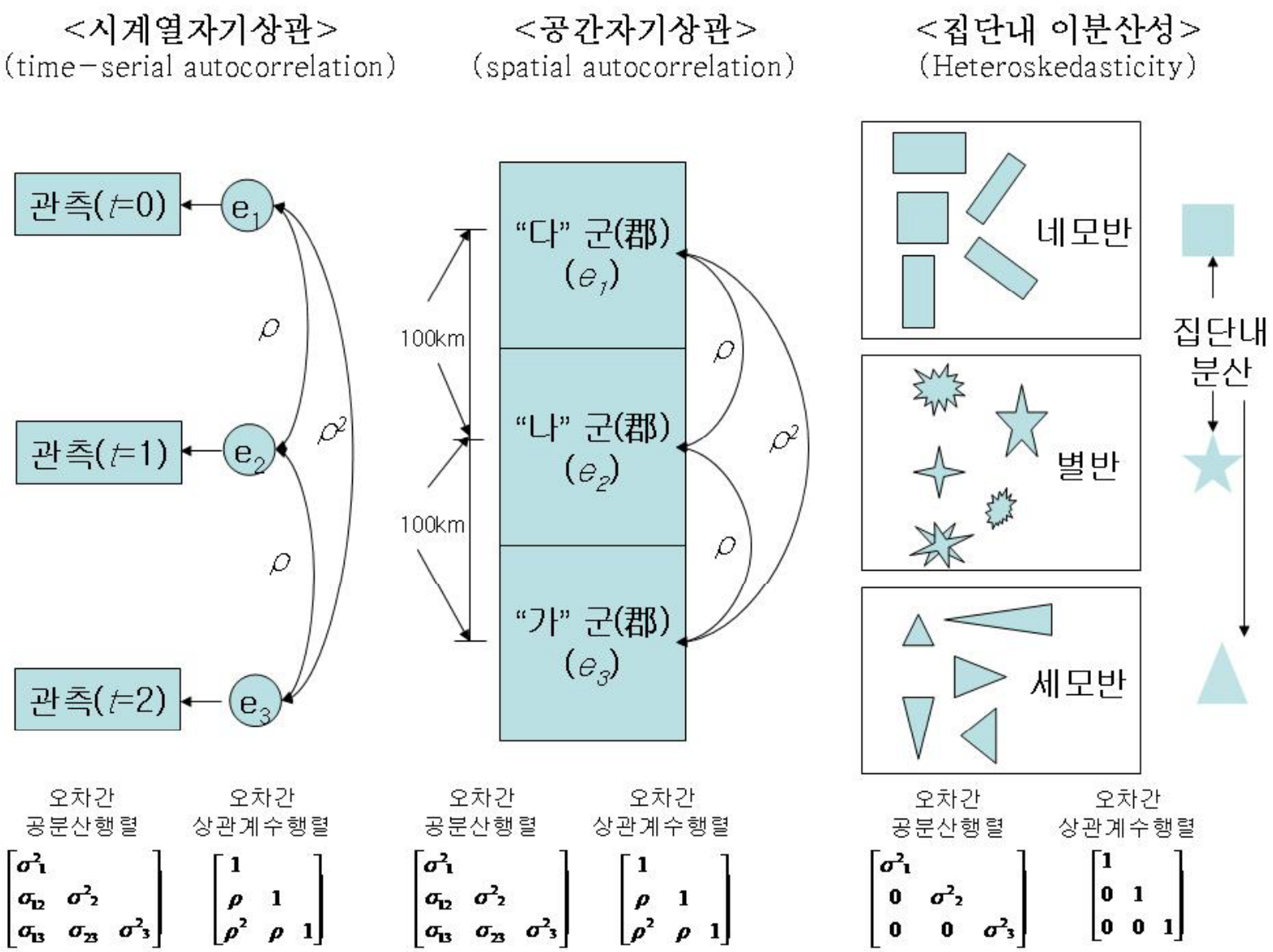
  예) 학생성적= 학생들 공부시간 + 교사 강의 스타일(주입식 대 토론식)
  개인수준 집단 수준

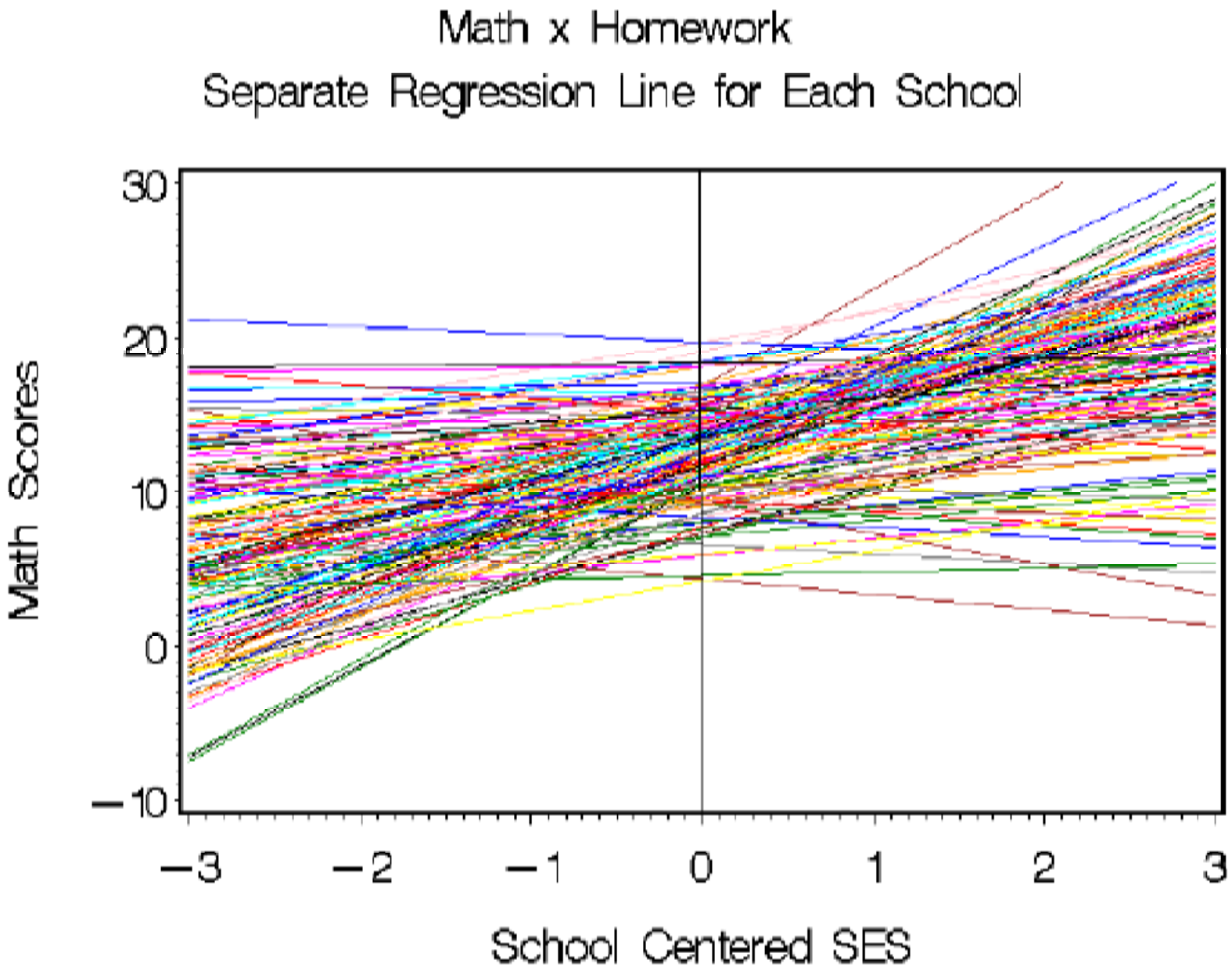
- 일반적인 회귀모형(OLS 혹은 로짓/프로빗 모형)들이 오직 하나의 오차항(error term)을 가정하는 반면, 다층모형에서는 측정수준별로 여러 개의 오차항들(error terms)을 갖는다.
  OLS: R-square, 다층모형 R = R(공부시간) + R(강의 스타일)

- 다층모형의 모수추정은 모두 반복적 수렴(iterative convergence) 과정을 통해서 추정된다. 쉽게 말하면 다층모형의 추정은 일반적인 회귀모형인 OLS나 로짓/프로빗 모형과는 달리 연구자들에 따라, 또한 통계분석 패키지에 따라 약간씩 상이할 수 있다.

다층모형은 일반 회귀모형을 보다 정교화시킨 모형!

자료의 다층적 구조와 오차 간 이분산성 구조




Math x Homework
Separate Regression Line for Each School
Math Scores
30
20
10
0
−10
−3
−2
−1
0
1
2
3
School Centered SES

아래와 같은 가상적인 자료를 보면 자료분석과 해석 시 개인차반영이 얼마나 중요한지 쉽게 알 수 있음.

학교 A	학기말성적	90	80	70	60	50	40	30
	입학시험성적	100	95	90	85	80	75	70
학교 B	학기말성적	89	88	78	68	58	48	38
	입학시험성적	65	60	55	50	45	40	35

만약 학교에 따른 차이를 무시할 경우 r_{AB}=.336 (p=n.s., n=14)로 나타나 학기말 성적과 입학시험성적은 아무런 관련이 없다고 결론내릴 수 있다.

그러나 학교별로 상관계수를 구하면 학교 A의 경우 r_A=1.00 (p<.001, n=7), 학교 B의 경우는 r_B=.991 (p<.001, n=7)로 학기말 성적과 입학시험 성적은 매우 밀접한 상관관계를 갖고 있음을 알 수 있다.

이것이 하위집단을 무시해서 발생한 전형적인 '집합의 오류'이다.

다층모형의 구성과 종류

다층모형의 구성과 추정은 학자들에 따라서 조금씩 다르지만, 흔히 다음의 세 단계로 구성되어 있다

1. 측정수준별로 개별방정식들을 구성

2. 개별방정식들을 하나의 통합방정식으로 구성

3. 통합방정식 내에서 상위 및 하위 수준의 효과를 통계적으로 구분

<종류>

1. 변화모형: 패널 자료를 이용한 다층모형으로 흔히 개체성장모형으로도 불린다

2. 배속모형: 개인의 관측 값이 위계적으로 배속된 집단에 따라 군집(群集, cluster)을 이루는 경우

변화모형(Growth Curve Model)

- Purpose is to model change over time
- Linear or nonlinear models possible
- Variability in change over time by modeling individual growth curves
- Variability in initial or average levels
- Predictors can be used to account for variability
- Two general approaches
    - Hierarchical linear models (HLM)
    - Structural equation models (SEM)

-- 자료가 시계열적(serially)으로 수집되는 패널 데이터를 분석하는 방법으로는 자기회귀분석법(auto-regression analysis)과 개체성장모형(individual growth curve model)이 있음.

-- 이 두 통계적 분석 모델은 종단적 패널 데이터 분석에서 모두 유용하게 쓰일 수 있다. 두 방식에서 어느 것이 더 우월한 방법이냐는 질문에 대한 답은 없음 (Bollen & Curran, 2004).

-- 자귀회귀분석에서는 영향력의 시차(time-lag)을 구체적으로 분석 모델에(예: 1st. lag, 2nd. lag 등) 집어넣을 수 있는 장점이 있음.

-- 성장모형에서는 시간의 흐름에 따른 개인의 변화 차이를 모델에 변인화 해 넣을 수 있는 이점이 있음.

-- 역으로 각 방법의 장점이 다른 방법의 단점이 된다. 따라서 바로 직전 시점(t-1)의 영향이 어떻게 작용하는가에 더 관심이 있는 연구문제는 자기회귀 방법을, 그보다는 시간 경과에 따른 개인의 변화 차이에 더 관심이 있을 경우에는 개체 성장모형을 사용.

변화성장모형(Growth Curve Model)의 확산 이유?

- availability of longitudinal data
- emphasis on individual differences
- accessibility of SEM software

HLM Approach to Growth Curves

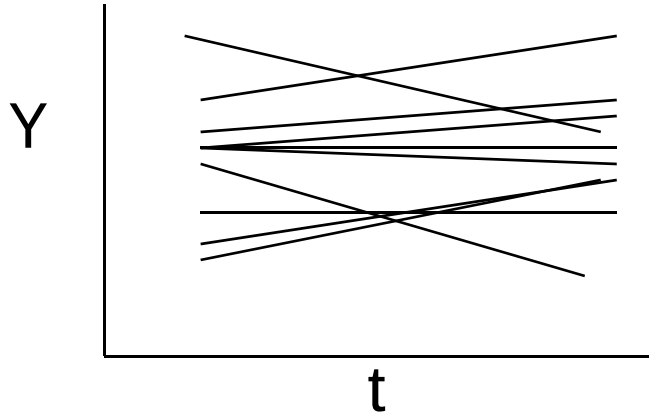
Conceptualization

- Two levels: within individual and between individual

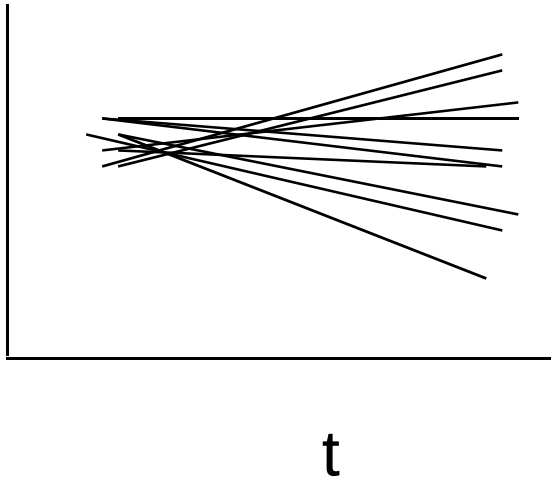
- Regression equation for each level

Example Growth Curves

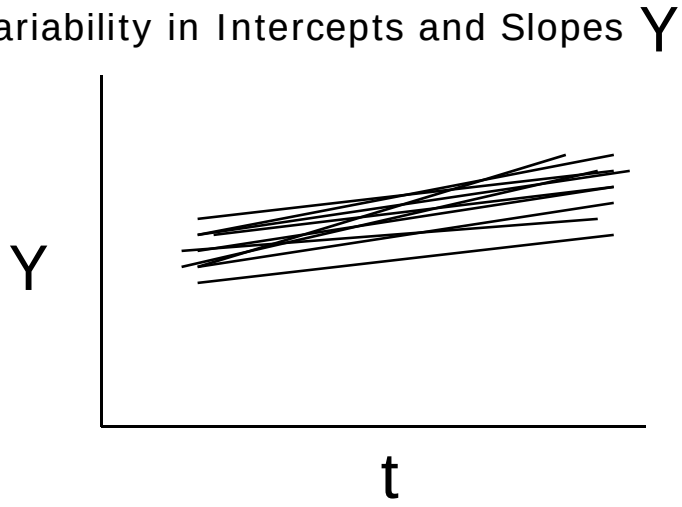
High Variability in Intercepts and Slopes



Low Variability in Intercepts and High Variability in Slopes



Low Variability in Intercepts and Slopes



HLM Approach to Growth Curves

Level 1: Time Level

$$Y_{ti} = \beta_{0i} + \beta_{1i}x_{ti} + r_{ti}$$

- Examines change in the dependent variable as a function of time for each individual

- Intercepts and slopes obtained for each individual

- Intercept is initial or average value of the dependent variable for a given individual (depending on coding of time variable)

- Slope describes linear increase or decrease in the dependent variable over time of a given individual

- With predictors, intercepts and slopes represent adjusted means and slopes

HLM Approach to Growth Curves

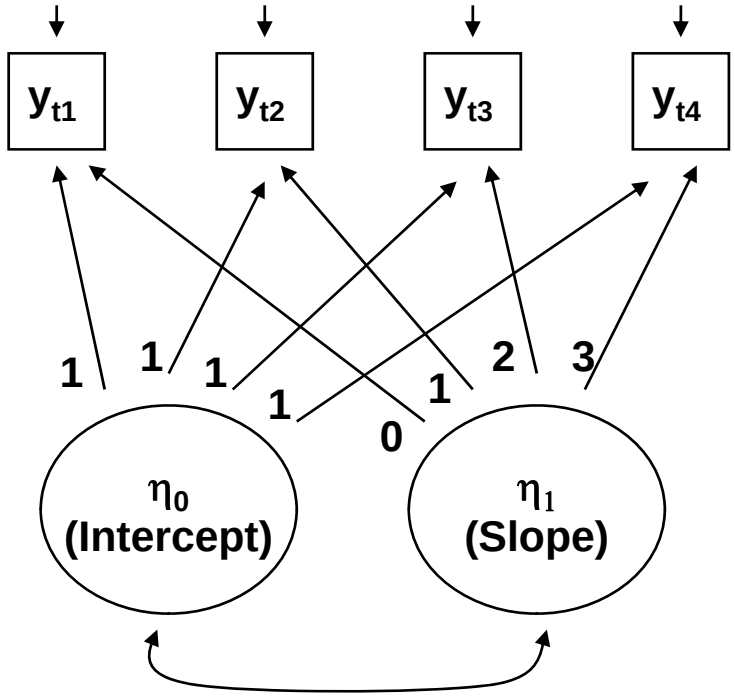
Level 2: Individual Level

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_{1i} + U_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} z_{1i} + U_{1j}$$

- **Intercepts and slopes obtained from Level 1 serve as dependent variables**
- **With no predictors, Level 2 intercept represents average of intercepts or slopes from Level 1**
- **With no predictors, Level 2 residual provides information about variance of intercepts or slopes across individuals**
- **Can incorporate predictors measured at the individual level (gender, income, etc.)**
- **Predictors explain variation in intercepts or slopes across individuals**

SEM Approach to Growth Curves

Example of a latent growth curve model with four time points



분석사례: 인터넷 일탈행위 연구

- 분석자료:

-- 한국청소년정책연구원에서 진행한 한국청소년패널조사(KYPS)의 초등학교 4학년 패널의 1, 2, 3 차 (2004년, 2005년, 2006년) 자료

-- 데이터는 2004년 1월 기준 교육통계연보를 표집틀로 하여 제주도를 제외한 전국 초등학생들을 대상으로 층화 다단계 집락표집(stratified multi-stage cluster sampling)을 통해 2,949명의 학생과 해당 학생의 부모를 추출하였고 설문조사는 2004년부터 1년 간격으로 진행되었음.

결과변인: 인터넷 일탈행동

-- 인터넷 게시판 등에 고의로 거짓 내용을 퍼뜨리기
-- 인터넷에서 불법소프트웨어 다운 받아 사용하기
-- 다른 사람의 인터넷 ID/주민등록번호를 허락 받지 않고 사용하기
-- 채팅하면서 성, 나이 속이기
-- 다른 사람의 컴퓨터/웹사이트 해킹하기
-- 채팅/게시판에서 상대방에게 욕설/폭언하기'등이다.

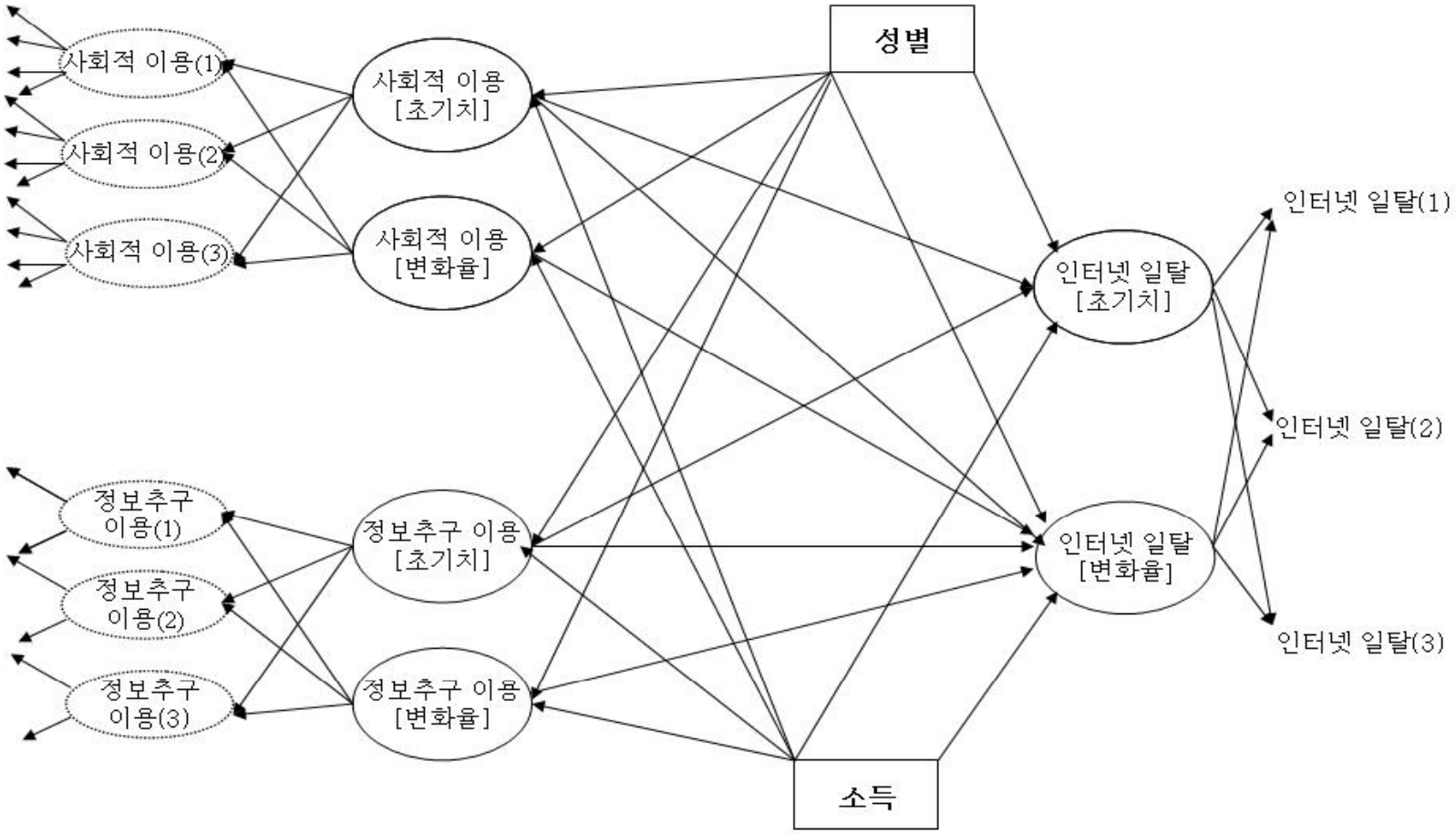
예측변인

1. 시간변동 예측변인: 인터넷 이용유형

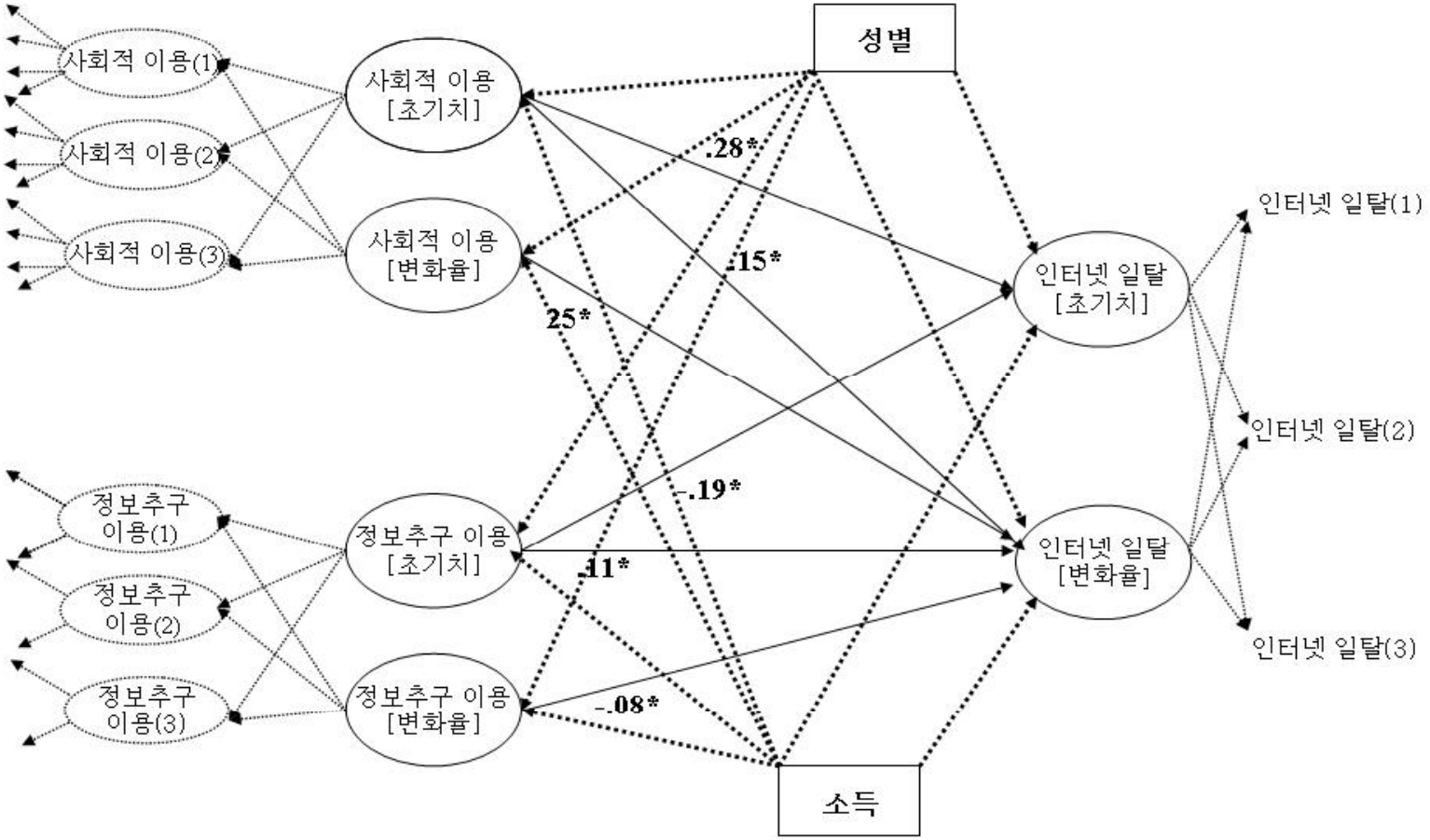
1) 사회적 이용
   -- 동호회/카페/커뮤니티 활동
   -- 전자우편(E-mail) 이용
   -- 채팅하기/메신저 사용
   -- 게시판 활동
2) 정보추구 이용
   -- 공부 및 학습관련 정보검색, 열람
   -- 기타 정보검색, 열람

2. 시간 무변동 예측변인: 성별, 가구 소득

분석연구 모형


사회적 이용(1)
사회적 이용(2)
사회적 이용(3)
정보추구 이용(1)
정보추구 이용(2)
정보추구 이용(3)
사회적 이용 [초기치]
사회적 이용 [변화율]
정보추구 이용 [초기치]
정보추구 이용 [변화율]
성별
소득
인터넷 일탈 [초기치]
인터넷 일탈 [변화율]
인터넷 일탈(1)
인터넷 일탈(2)
인터넷 일탈(3)

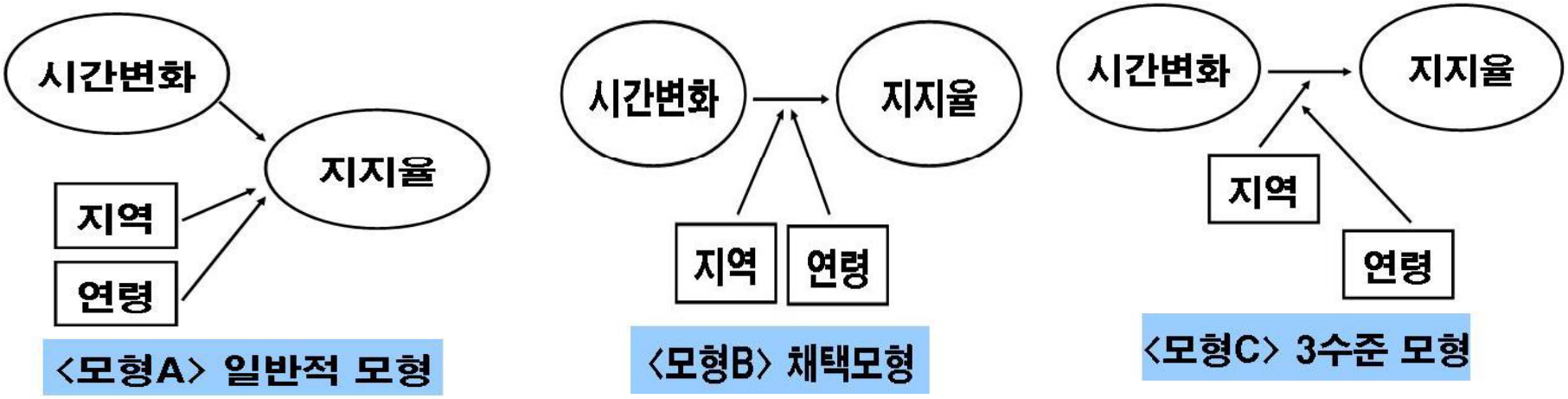
분석결과



RMSEA=.08 GFI= .86

배속모형(Embedded Model)

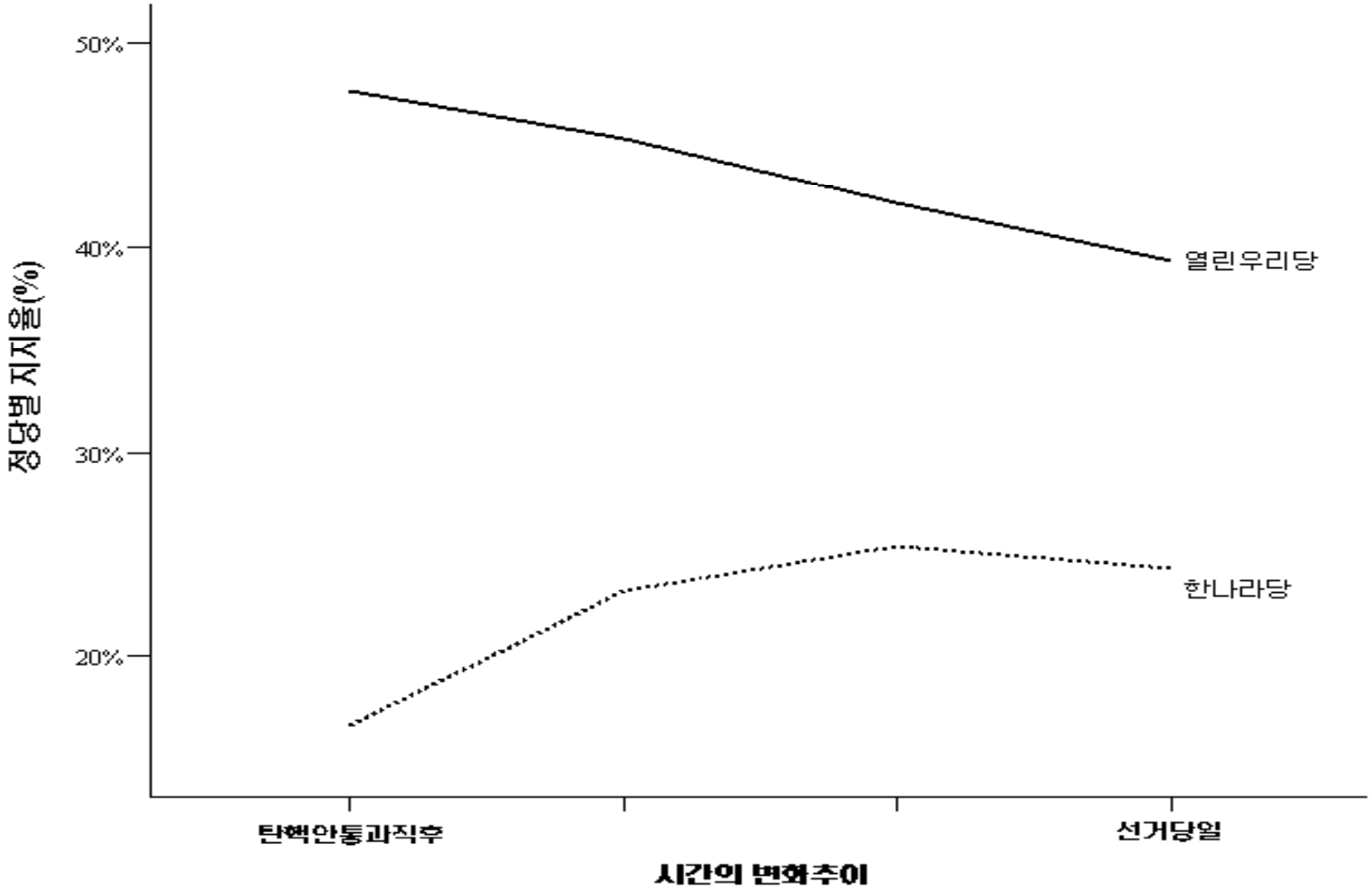
‘탄핵’의 정치적 효과와 지속시기: 계층적 선형모형을 중심으로



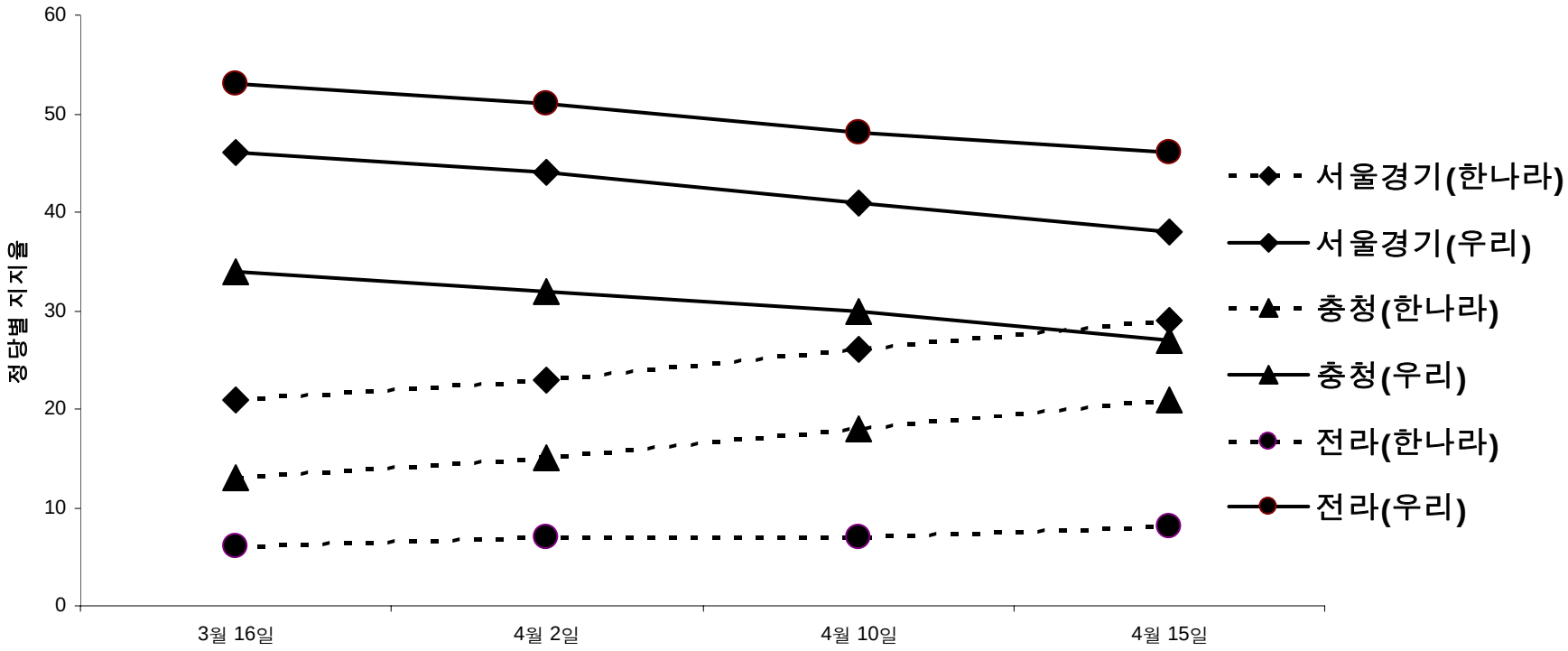
- 지역을 개인차를 반영하는 변수로 간주함

- 약 30여년간 유권자의 거주지와 지지후보의 거주지가 연결되어서 사용되었다는 점에서 개인차로 해석될 수 있다

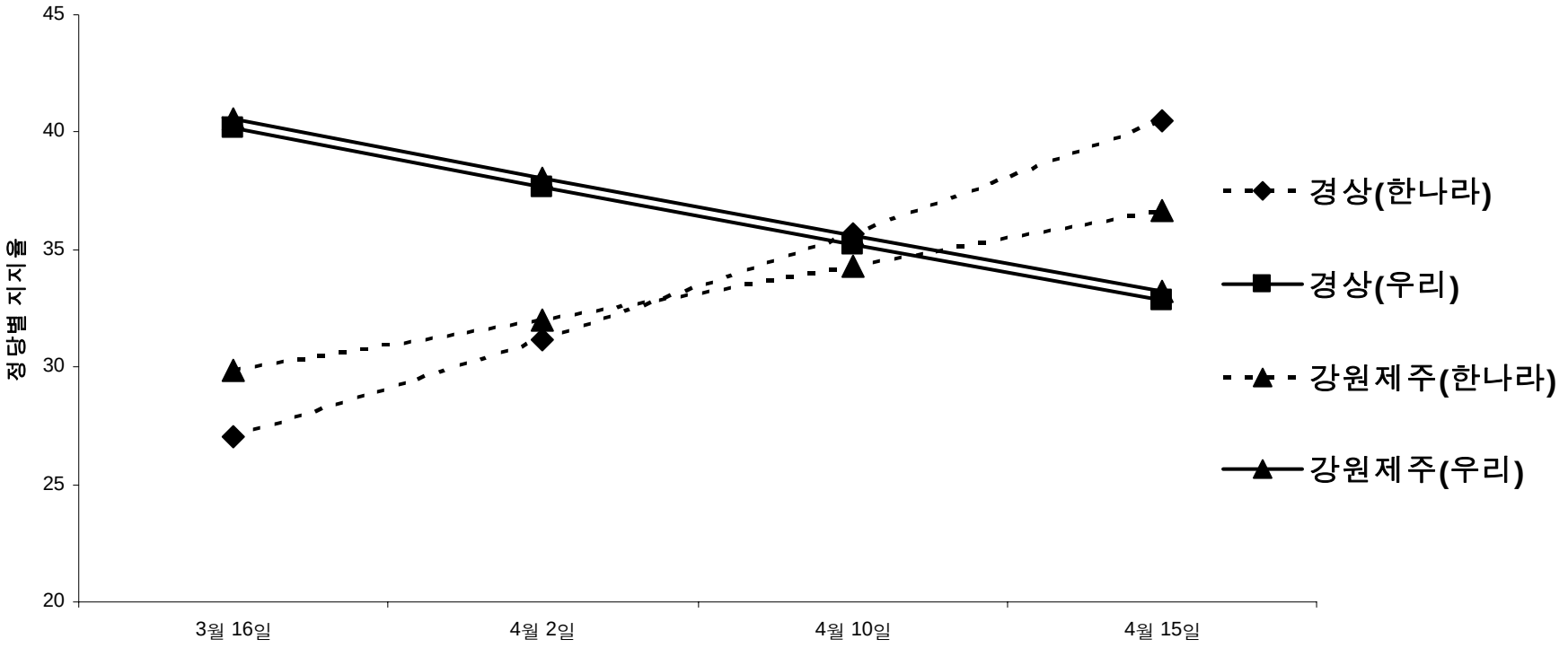
조사응답자들의 지지정당 선택변화



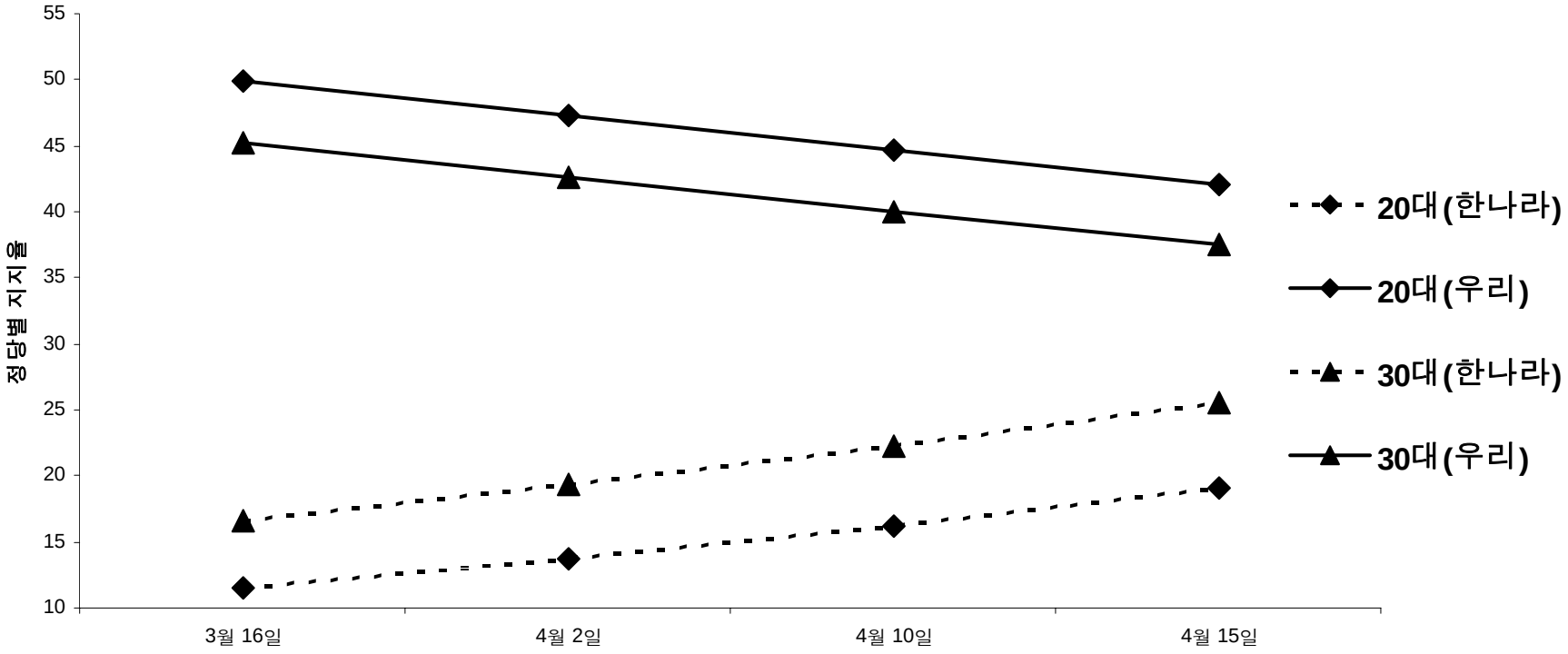
서울·경기, 충청도, 전라도 지역의 양당지지율 변화 추이



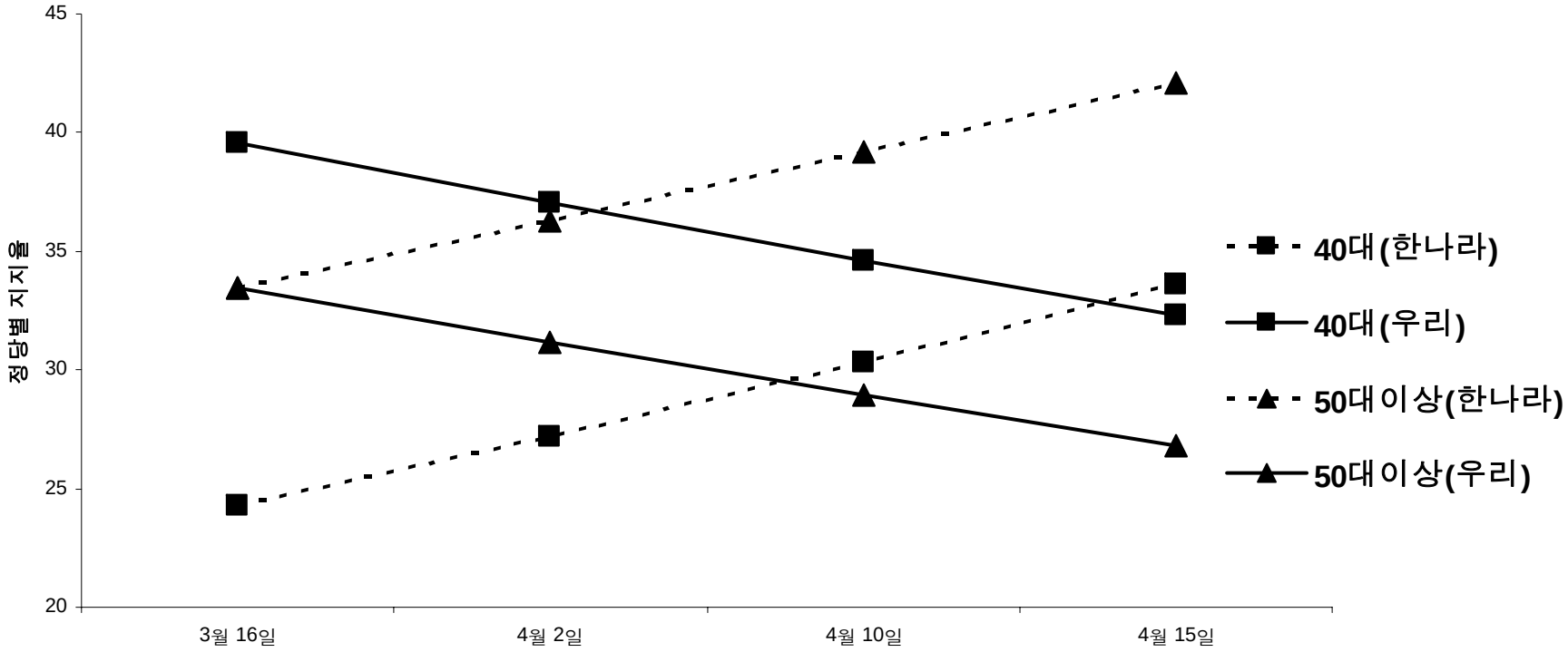
경상도, 강원·제주 지역의 양당 지지율 변화 추이



20·30대의 양당 지지율의 변화 추이



40·50대의 양당 지지율 변화추이

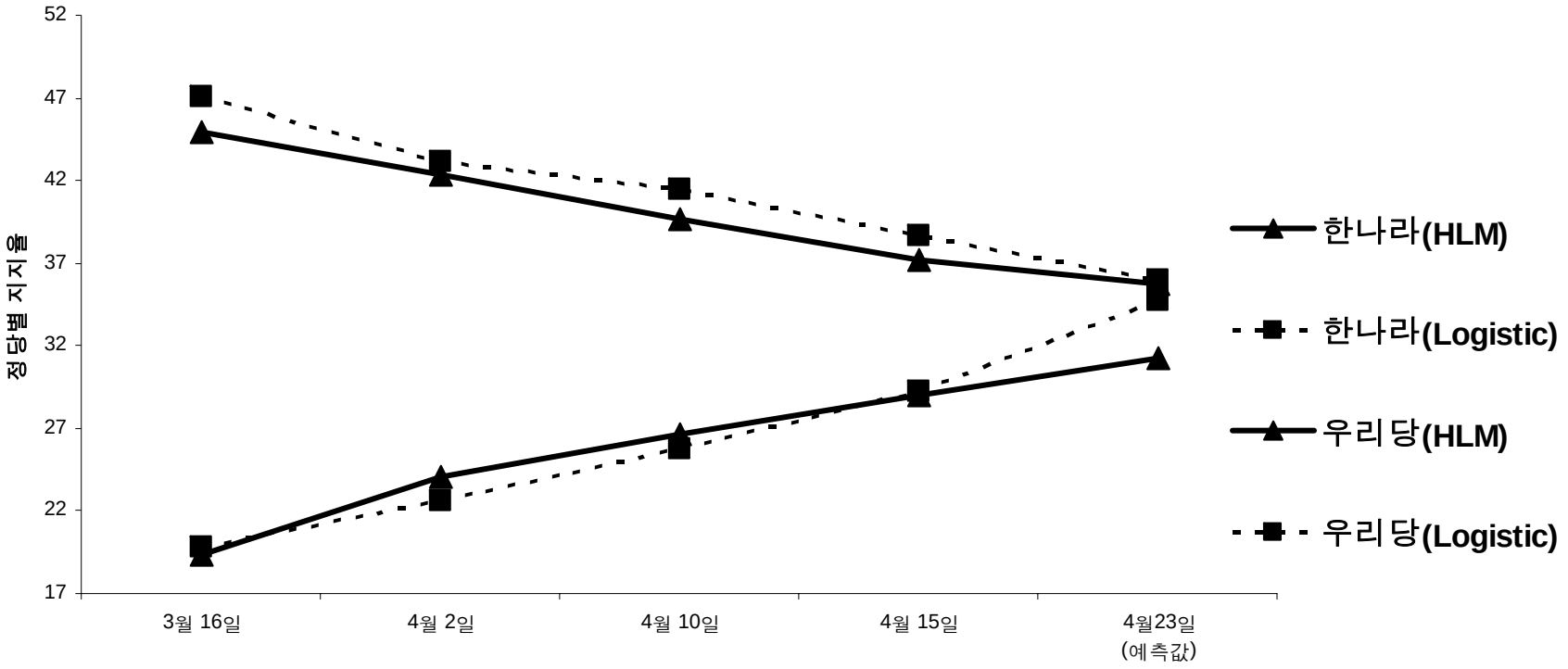


사시점별 일반 선형모형과 계층적 선형모형의 정당지지율 예측 값 비교

	한나라당		열린우리당	
	HLM (모형 3)	로지스틱회귀 모형	HLM (모형 2)	로지스틱회귀 모형
3월 16일	19.3	19.8	44.9	47.1
4월 2일	24.1	22.6	42.3	43.1
4월 10일	26.7	25.7	39.7	41.4
4월 15일	29.0	29.2	37.2	38.7
4월 23일 (예측값)	31.3	34.7	35.7	36.0

- 4월 23일 한국갤럽의 조사결과 한나라당에 대한 지지율은 27.9%,
- 열린우리당의 지지율은 35.3%.

계층적 선형모형과 일반 선형모형의 지지율 변화비교



다층모형, 과연 필요한가?

• 다층모형의 이용을 둘러싼 가장 큰 논란은 복잡성이다. 보통의 사회과학 연구들은 가변효과에 대해 그다지 많은 관심을 기울이지 않는다. 사회과학도들이 관심을 갖는 것은 고정요인에 의한 고정효과이며, 오차(error)로 처리되는 오차항들이 사회과학적 지식의 증진을 위해 얼마나 많은 기여를 할지는 미지수다.

• 다음으로 생각할 수 있는 문제는 배속모형에서의 집단수준의 표본사례수와 집단의 특성이다. 변화모형과는 달리 배속모형은 집단내에 배속되어 있는 개인들의 수가 천차만별일 경우가 많다.

• 또 자주 언급되는 논란은 모형추정방법이다. 대부분의 배속모형처럼 자료의 구조가 비균형설계(unbalanced design)를 따르면서 집단내 개인들의 사례수가 충분히 보장되지 않을 경우, 모형추정방법에 따라 결과가 매우 상이한 것으로 알려져 있다.

• 마지막 비판은 다층모형의 과용(過用)이다. 최근 다층모형이 인기를 얻으면서, 다층모형이 지나치게 과용되고 남용되고 있는 것이 사실이다. 회귀분석 가정들의 위반이라는 점에서 다층모형은 좋은 대안을 제공해줄 수 있다. 많은 경우 OLS의 결과와 비슷한 회귀계수가 나타나는데 유의도 수준에서만 차이가 난다.