Journal of the Korean Data & Information Science Society
2017, **28**(1), 1–10

http://dx.doi.org/10.7465/jkdi.2017.28.1.1
한국데이터정보과학회지

경시적 자료를 이용한 아동 학업성취도 분석[†]

이나은[1] · 허집[2]

[12]덕성여자대학교 정보통계학과

접수 2016년 12월 21일, 수정 2017년 1월 4일, 게재확정 2017년 1월 5일

요 약

경시적 자료를 이용한 아동 학업성취도에 영향을 주는 요인을 찾기 위한 기존의 분석들은 각 아동의 반복 측정된 자료들이 독립이라고 가정한 모형을 주로 이용하였다. 본 연구에서는 기존 연구들에서 고려한 아동 학업성취도에 영향을 주는 변수들을 선택하여 반복 측정된 경시적 자료의 종속성을 고려한 고정효과와 임의효과를 포함하는 선형혼합모형으로 분석하여 아동 학업성취도에 영향을 주는 변수들은 무엇인지, 각 아동의 특성들이 반영되는 임의절편과 임의기울기가 있는지를 파악하는 것이 연구의 목적이다. 본 연구에 사용된 자료는 한국복지패널 1, 4, 7차 부가조사 중에서 아동용 설문문항에 대한 자료이고, 국어, 영어와 수학의 학업성취도 점수의 합을 아동 학업성취도로 한다. 선형혼합모형을 이용한 분석 시에 다중공선성의 검토와 결측치의 특성을 파악하고 적절한 오차의 상관행렬을 선택한다.

주요용어: 고정효과, 다중공선성, 상관행렬, 선형혼합모형, 임의결측, 임의효과.

1. 서론

교육열을 표현한 말들 중 '맹모참천지교'만큼 널리 쓰이고 익숙한 말은 없을 것이다. 대학입시 하나의 목적에 맞춰져 있는 우리나라의 교육열은 세계 어느 나라와 비교하여도 낮지 않을 정도로 우리나라 발전의 원동력이면서도 병폐이기도 하다. 이러한 교육열을 만족시키기에는 공교육 지원의 한계로 인해 사교육 시장은 교육열의 과열에 비례하여 커져만 가고 있다. 부모의 사회적·경제적 지위와 관련이 있는 사교육의 정도가 과연 아동 및 청소년들의 학업성취도에 영향을 주고 있는지 의문을 가져볼 필요가 있다.

이러한 학업성취도에 대하여 경시적 자료 (longitudinal data)를 이용한 연구가 다양하게 진행되었다. 다중회귀분석을 실시한 Kim (2010)은 저소득층은 사교육비가 학업성취도에 유의한 영향이 없었지만 고소득층에서 사교육비는 매개변수로 학업성취도에 유의한 영향을 줌을 보였다. Kim과 Lee (2015)는 구조모형 분석으로 부모의 성취 지향적 양육태도가 학업성취도에 부정적인 영향을 줌을 보였다. Yoo와 Park (2015)은 구조방정식을 이용하여 부모의 사회적·경제적 지위가 높을수록 사교육에 대한 지출이 많아지는 경향이 있지만 사교육 지출이 학업성취도에 미치는 영향은 유의하지 않다는 것을 보였다.

경시적 자료가 아닌 경우의 과거 연구들을 살펴보면, 고소득층에서 부모의 양육행동은 학업성취도에 유의함을 보였다. 경로분석을 이용한 Seol과 Jung (2013)의 연구에서는 부모의 학습관여가 아동의 학

[†] 본 연구는 덕성여자대학교 2015년도 교내연구비 지원에 의해 수행되었음.
[1] (01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과, 석사과정.
[2] 교신저자: (01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과, 교수.
E-mail: jhuh@duksung.ac.kr.

교 소속감을 매개로 하여 학업성취에 영향을 주었다. 경기교육종단연구 2차 년도 자료를 사용해 다중회귀분석을 한 Chung과 Jeong (2015)의 연구에서는 부모의 교육적 관여 수준이 학업성취도에 긍정적인 영향을 주었다.

지금까지 아동 학업성취도에 대한 기존 연구들은 주로 각 아동들의 독립적 반복 측정된 자료들이라는 가정 하에 다중회귀분석이나 구조방정식 모형을 이용해 분석하였다. 본 연구는 한국복지패널 부가자료인 1, 4, 7차 년도의 자료를 사용하여 각 아동의 반복 측정된 자료의 종속성을 가정하고, 기존 연구 결과에서 아동 학업성취도에 영향을 주거나 아동 학업성취도와 관련 있는 주요 설명변수들을 사용해 아동 학업성취도에 관한 분석을 선형혼합모형 (linear mixed model)으로 하고자 한다. 반복 측정된 경시적 자료를 이용한 선형혼합모형과 일반화선형혼합모형에 대한 최근의 연구로는 Lim과 Baek (2012), Jo와 Chang (2013), Shim 등 (2013), Kim 등 (2014), Jeon과 Lee (2014)와 Zhang과 Baek (2015) 등이 있다.

2절에서는 자료 및 변수의 설명과 변수간의 특성에 대하여 커널추정법을 이용하여 알아보고자 한다. 또한, 반복 측정 자료의 종속성을 파악하고자 상관분석을 실시한다. 3절에서는 상관행렬과 선형혼합모형에 대하여 알아보고, 모형 적합 전에 다중공선성의 존재 여부와 결측치의 특성을 조사하고자 한다. 4절에서는 최종모형을 선택하여 아동 학업성취도에 영향을 미치는 설명변수와 임의효과 (random effect)가 있는지 파악하고자 한다.

2. 자료의 설명과 특성

2.1. 자료 및 변수의 설명

한국복지패널 자료는 한국보건사회연구원과 서울대학교 사회복지연구소가 공동으로 생산하였으며 매년 1회씩 조사되고 있다. 아동용, 복지인식부가조사용, 장애인부가조사용은 3년마다 한 번씩 돌아가며 조사된다. 아동용 조사표의 경우 2006년을 기준으로 3년마다 조사되며 조사시점에 초등학교 4∼6학년에 재학 중인 아동이 조사대상이다. 아동 부가조사는 지금까지 2006년, 2009년, 2012년과 2015년에 조사되었지만, 2015년 조사에서는 조사 대상이 과거 조사 대상자들이 아닌 새로운 초등학생 4학년을 대상으로 한 조사이기에 분석대상에서 제외하였다. 각 연도별 조사 대상자 수는 662명, 515명, 436명이며 성별, 학년별 조사 대상자 수는 Table 2.1에 주어졌다. 학년은 2006년 조사 당시의 학년을 기준으로 표시하였다.

Table 2.1 Number of data for each year, gender, grade and parents' educational background

year		2006	2009	2012
gender	male	340	257	219
	female	322	258	217
grade in 2006	4^{th} grade	220	168	136
	5^{th} grade	213	171	152
	6^{th} grade	229	176	148
parents' educational background	low	431	343	294
	high	231	172	142
total		662	515	436

한국복지패널 아동부가조사는 국어, 영어, 수학의 학교성적을 각각 5점 척도로 나누었으며, 이 문항들의 각 아동의 설문조사 결과의 합으로 아동 학업성취도 (academic achievement)로 사용하였다. 본 연구에서는 아동 학업성취도를 반응변수인 y로 두고, 양적 설명변수들을 다음과 같이 월평균 사교육비 (private education expenditure)를 x_1, 부부 및 가족갈등 (marital and family conflict)을 x_2, 부정

적 양육행동 (negative parenting behavior)을 x_3, 자아존중감 (self-esteem)을 x_4, 주택의 면적 기준 (housing area)을 x_5라 하였다. 월평균 사교육비의 단위는 십만원이다. 각 설명변수들은 설문조사의 관련된 항목들의 점수의 합으로 하였다.

아래 Table 2.2는 각 설명변수들과 반응변수의 연도별 표본평균과 표본표준편차이다. 질적 설명변수인 부모 교육수준 (parents' educational background)은 부모 중 한 사람이라도 전문대를 졸업한 고학력인 경우와 그렇지 않은 저학력인 경우로 나누어 변수화 하였다. 부모 교육수준별 자료의 빈도는 Table 2.1에 제시하였다.

Table 2.2 Sample means (SM) and sample standard deviations (SSD) of response and exploratory variables

year	2006		2009		2012	
variable	SM	SSD	SM	SSD	SM	SSD
y	10.48	2.59	9.50	2.93	8.81	2.66
x_1	3.28	3.11	3.91	3.95	3.31	4.17
x_2	9.72	2.37	9.78	2.30	9.88	2.08
x_3	8.84	1.77	9.33	2.52	9.10	2.51
x_4	38.64	5.62	38.40	6.27	37.95	6.07
x_5	2.04	0.79	2.14	0.83	2.29	0.87

Table 2.3은 각 연도별 아동 학업성취도에 대한 성별과 학년별로 표본평균과 표본표준편차를 보여주고 있다. 성별, 학년별 아동 학업성취도는 조사가 진행될수록 표본평균이 낮아지는 추세를 보이고 있다. 큰 차이를 보이고 있지는 않지만, 남학생의 학업성취도가 여학생의 학업성취도에 비해 높은 경향이 있다. 2006년의 학업성취도에서는 학년별로 큰 차이를 보이지 않고 있지만, 2006년도의 4학년의 경우가 타 학년의 경우보다 2009년과 2012년의 학업성취도가 높아 보인다.

Table 2.3 Sample means (SM) and sample standard deviations (SSD) of child academy achievement for each year, grade and gender

year		2006		2009		2012	
SM / SSD		SM	SSD	SM	SSD	SM	SSD
gender	male	10.500	2.683	9.623	2.847	8.909	2.625
	female	10.460	2.494	9.385	3.006	8.705	2.695
grade in 2006	4^{th} grade	10.441	2.617	9.826	2.877	9.007	2.696
	5^{th} grade	10.507	2.665	9.421	2.988	8.612	2.602
	6^{th} grade	10.493	2.505	9.278	2.904	8.824	2.687

2.2. 변수들의 특성

양적 설명변수들과 반응변수간의 전반적인 회귀함수 관계를 파악해보고자 커널함수를 이용한 국소선형추정량 (local linear estimator)을 이용하였다. 이때 쓰인 커널함수는 Epanechnikov 커널을 이용하였고, 정밀한 함수 관계를 추정하고자 하는 것이 아니라 전반적인 함수의 특성을 파악하기 위한 것이기에 대략적으로 각 설명변수의 범위의 1/10을 띠폭 (bandwidth)으로 사용하였다. 커널추정량에 대한 자세한 내용은 Fan과 Gijbels (1996)를 참조하기 바란다.

Figure 2.1은 각 설명변수와 반응변수와의 국소선형추정량의 결과를 보여주고 있다. 월평균 사교육비 (x_1), 자아존중감 (x_4)와 주택의 면적 기준 (x_5)이 학업성취도에 긍정적 영향을 주는 경향이 있어 보이고, 부정적 양육행동 (x_3)은 오른쪽 자료의 희박성을 감안한다면 전반적으로 부정적 영향을 주는 것처럼 보인다. 부부 및 가족갈등 (x_2)은 아동의 학업성취도에 영향을 주지 않는 것처럼 보인다.

한편, 반복 측정된 경시적 자료의 반응변수의 자료들의 종속성의 존재여부를 보기 위하여 아래의 Figure 2.2에서 각 반복 측정 자료들 사이의 표본상관계수와 모상관계수가 0인지 아닌지 가설검정에 대한 유의확률을 보여주고 있다. 이 결과를 보면 각 반복 측정 자료들은 양의 상관관계를 가지고 그들 사이의 모상관계수가 0이 아닌 것처럼 보이며, 조사년도의 차가 클수록 표본상관계수가 작아지는 경향을 보이고 있다. 즉, 1차와 2차 혹은 2차와 3차보다는 1차와 3차 사이의 표본상관계수가 작은 경향을 보이고 있다. 세 개의 히스토그램은 각 연도별 아동의 학업성취도에 대한 것이다.

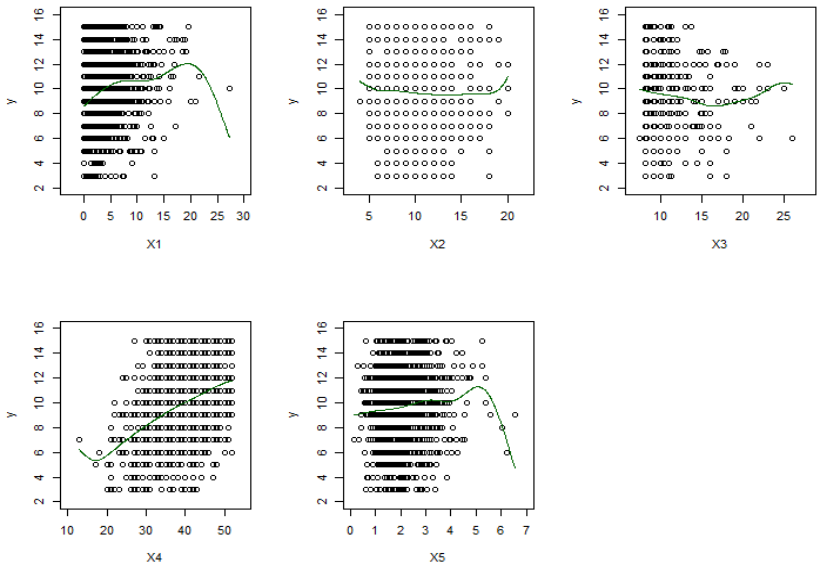

Figure 2.1 Local linear estimates of regression functions for each exploratory variable

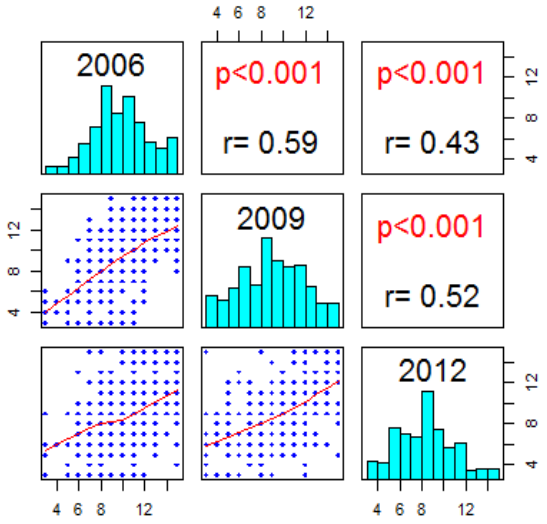

Figure 2.2 Sample correlations with p-values and histograms of child academic achievement for among years

3. 선형혼합모형의 적합

3.1. 선형혼합모형

관측 대상인 개체의 수를 m이라 할 때, i번째 개체의 반복 측정 자료의 수를 n_i라 하자. 설명변수의 수가 p개인 경우에, i번째 개체의 j번째 종속변수의 관측치를 y_{ij}라 하고 그 때의 설명변수들을 x_{1ij}, x_{2ij}, $\cdots$, x_{pij}라 하면 선형혼합모형은 다음과 같이 서술된다.

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \varepsilon_{ij},\ i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, n_i, \tag{3.1}$$

여기서 $\boldsymbol{x}_{ij} = (1, x_{1ij}, x_{2ij}, \cdots, x_{pij})^T$는 $(p+1)$차 벡터이고, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$는 고정효과 (fixed effects)로서 $\boldsymbol{x}_{ij}$에 관련된 모수벡터이다. 다음 $p \geq q$을 만족하는 양의 정수 q에 대하여, $(q+1)$차 벡터인 $\boldsymbol{z}_{ij} = (1, z_{1ij}, z_{2ij}, \cdots, z_{qij})^T$는 $\boldsymbol{x}_{ij}$벡터의 일부분으로 구성되고 i번째 개체의 j번째 반복측정에 대한 임의효과에 관한 보조벡터이다. 확률벡터 $\boldsymbol{b}_i = (b_{0i}, b_{1i,} \cdots, b_{qi})^T$는 임의효과이며 ε_{ij}은 오차항이다. 임의효과와 오차항들은 서로 독립이라 가정한다. 서로 다른 개체들의 오차항끼리는 서로 독립이며, 각 개체의 반복 측정에 대한 오차항들은 독립이 아니어서 공분산들을 가진다고 가정한다. 즉, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{in_i})^T$라 하면 $\boldsymbol{\varepsilon}_i$와 $\boldsymbol{b}_i$의 확률분포는 각각 n_i차와 m차 다변량정규분포로 다음과 같이

$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}_{n_i}, \Sigma_i) \text{와 } \ \boldsymbol{b}_i \sim N_q(\mathbf{0}_q, D) \tag{3.2}$$

이고 $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_m, \boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m$은 서로독립이라 가정한다. 식 (3.2)에서 자연수 k에 대하여 N_k와 $\mathbf{0}_k$는 k차원 정규분포와 0벡터이다. 행렬 Σ_i와 D는 각각 $\boldsymbol{\varepsilon}_i$와 $\boldsymbol{b}_i$의 분산-공분산행렬이다. 식 (3.1)에서 임의효과가 없는 경우에 선형혼합모형은 일반선형모형 (general linear model)이 된다. 선형혼합모형에 대한 자세한 내용은 Verbeke와 Molenberghs (2009)를 참고하길 바란다.

식 (3.2)의 오차 $\boldsymbol{\varepsilon}_i$의 분산-공분산행렬 Σ_i를 이루는 상관행렬 (correlation matrix)로 흔히 쓰이는 것들 중에서 교환가능 (exchangeable)행렬은 한 개체 내의 반복 측정에 의한 오차들은 반복 측정 시차와 관련 없이 모두 같은 상관계수 ρ를 가질 때 이용된다. 즉, 임의의 서로 다른 j, j'에 대해 $corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho$를 만족한다. 일차자기상관 (first-order autocorrelation, AR(1))행렬은 임의의 서로 다른 j, j'에 대해 $corr(\varepsilon_{ij}, \varepsilon_{ij'}) = \rho^{|j-j'|}$을 만족하여 시차가 클수록 상관계수가 작아지는 상관행렬 모형이다. 특별한 상관계수 구조를 가지지 않을 경우에는 비구조적 (unstructured)행렬이 쓰이고 있다. 비구조적행렬은 상관행렬에 특정한 형태를 가정하지 않은 모형으로 $n_i(n_i-1)/2$개의 상관계수가 모두 모수가 된다. 이들 상관행렬에 대한 자세한 내용이나 그 외 사용되는 상관행렬에 대한 것은 Diggle 등 (2001)을 참고하길 바란다. 만약, 각 개체 내의 오차항들이 독립인 경우에는 상관행렬은 단위행렬이 되어 분산-공분산행렬 Σ_i들은 오차항의 분산들만으로 구성된 대각행렬이 된다. 각 개체별로 다음과 같이 $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$, $X_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots \boldsymbol{x}_{in_i})^T$, $Z_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \cdots \boldsymbol{z}_{in_i})^T$로 행렬과 벡터를 정의하면, 식 (3.1)을 다음과 같이 표기할 수 있다.

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{3.3}$$

여기서 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_m^T)^T$, $X = (X_1^T, X_2^T, \cdots, X_m^T)^T$, $Z = diag(Z_1, Z_2, \cdots, Z_m)$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, \cdots, \boldsymbol{b}_m^T)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \boldsymbol{\varepsilon}_2^T, \cdots, \boldsymbol{\varepsilon}_m^T)^T$이다.

3.2. 선형혼합모형의 적합

3.2.1. 다중공선성과 결측치

설명변수들간의 선형종속성인 다중공선성 (multicollinearity)의 존재 여부를 판단하기 위하여 분산팽창계수 (variance inflation factor; VIF)를 계산하여 Table 3.1에서 보여주고 있다. Montgomery 등

(2012)에 설명된 분산팽창계수는 $\text{VIF}_j = (1 - R_j)^{-1}$, $j = 1, \cdots, k$이고 k는 설명변수의 개수이다. 여기서 R_j^2는 설명변수 x_j를 반응변수로 간주하고 나머지 설명변수들과의 중회귀모형의 적합에서 얻어진 결정계수이다. 일반적으로 VIF가 10이상인 경우에 다중공선성이 있다고 판별한다. 따라서 Table 3.1의 결과에 의하면 다중공선성은 존재하지 않은 것으로 판단된다.

Table 3.1 Variance inflation factors (VIF) of exploratory variables

variable	x_1	x_2	x_3	x_4	x_5
VIF	1.11	1.03	1.06	1.05	1.11

각 개체의 반복 측정으로 얻어지는 경시적 자료는 다양한 이유에 의해 조사 되지 못하는 경우가 있어 결측치 (missing value)의 발생은 필연적이라 할 수 있다. 본 연구에서도 마찬가지로 Table 2.1이 보여주듯이, 한국복지패널 아동 부가조사에서 1차년도인 2006년의 조사 아동 수는 662명이었으나, 4차년도와 7차년도에서는 결측치가 발생하여 조사되어진 아동의 수는 각각 515명과 436명으로 줄어들었다. 결측치는 발생의 특성에 따라 무시할 수 있는 결측인 완전임의결측 (missing at completely random), 임의결측 (missing at random)과 무시할 수 없는 결측인 비임의결측 (missing not at random)으로 구분한다. 자세한 내용은 Rubin (1976)과 Laird (1988)를 참조하기 바란다.

Diggle 등 (2001)은 같은 관측 연도 내에서 결측 횟수별 신뢰구간들이 모두 공통 구간을 가지는 경우에 무시할 수 있는 임의결측으로 간주할 수 있다고 하였다. 본 한국복지패널의 결측 횟수별 및 연도별로 반응변수의 모평균에 대한 99% 신뢰구간을 Table 3.2에서 보여주고 있고, 같은 관측 연도 내에서 결측 횟수별 신뢰구간들이 모두 공통 구간을 가짐을 알 수 있기에 본 자료는 무시할 수 있는 임의결측 자료라 가정할 수 있다. 최근의 결측치에 대한 연구로는 Yoon과 Choi (2012)가 있다.

Table 3.2 Confidence intervals of mean of response variable for each number of missing value and year

year		2006	2009	2012
number of missing values	0	(9.934, 10.580)	(9.067, 9.802)	(8.478, 9.137)
	1	(10.376, 11.827)	(9.079, 10.693)	
	2	(10.259, 11.360)		

3.2.2. 선형혼합모형의 적합

설명변수들간의 다중공선성의 문제점이 없으며, 무시할 수 있는 결측인 임의결측을 가정하여 선형혼합모형 (3.3)는 McCulloch 등 (2008)에 의하여 다음과 같이

$$\widehat{\boldsymbol{\beta}} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}, \tag{3.4}$$

$$\widehat{\boldsymbol{b}}_i = D Z_i^T V_i^{-1} (\boldsymbol{y}_i - X_i \widehat{\boldsymbol{\beta}}) \tag{3.5}$$

을 이용하여 $\boldsymbol{\beta}$를 추정하고 $\boldsymbol{b}_i$를 예측한다. 이들은 각각 $\boldsymbol{\beta}$의 최량선형불편추정 (best linear unbiased estimator)과 $\boldsymbol{b}_i$의 최량선형불편예측 (best linear unbiased predictor)임이 알려져 있다. 여기서, $V_i = Z_i D Z_i^T + \Sigma_i$, $V = diag(V_1, V_2, \cdots, V_m)$이다. 위 추정과 예측 과정에서 분산-공분산행렬들의 추정법으로 최대우도법 또는 제한적 (restricted) 최대우도법이 흔히 이용된다. 식 (3.4)와 (3.5)에 쓰인 설명변수들은 Table 2.2의 양적 변수인 설명변수 5개와 다음과 같이 질적 변수들을 가변수화 한 4개를 포함하여 모두 9개이다. 설명변수들 중 질적 변수인 성별, 학년, 부모의 교육수준들을 다음과 같이 가변수화 하였다. 설명변수 x_6은 질적 변수인 성별을 가변수화 한 것이고, x_7과 x_8은 학년을 가변수화 한 것이다. 그리고 x_9은 부모의 교육수준의 가변수이다. 이들 가변수들은 다음과 같다.

$$x_6 = \begin{cases} 1, & \text{여학생} \\ 0, & \text{남학생} \end{cases}, \; x_7 = \begin{cases} 1, & \text{5학년} \\ 0, & \text{4학년, 6학년} \end{cases}, \; x_8 = \begin{cases} 1, & \text{6학년} \\ 0, & \text{4학년, 5학년} \end{cases}, \; x_9 = \begin{cases} 1, & \text{전문대졸 이상} \\ 0, & \text{그 외} \end{cases}.$$

위 식 (3.4)와 (3.5)를 계산하기 위하여 제한적 최대우도법을 선택하고 Lindstrom과 Bates (1988)에 의한 Newton-Raphson 반복법과 EM알고리즘으로 구현된 R 소프트웨어에서 제공하고 있는 함수 lme()를 사용하였다.

4. 모형의 선택 및 결론

아동 학업성취도에 영향을 주는 설명변수들이 어떤 것인지, 영향을 주는 변수들이 고정효과만 있는지, 임의효과도 있는지 등을 알아보고자 식 (3.3)의 선형혼합모형에서 최적의 모형을 찾고자 한다. 먼저 고정효과만 있는 모형을 고려하여 회귀계수를 추정하였다. 3.1절에서 설명하였듯이 오차의 상관행렬은 단위행렬, 교환가능행렬, AR(1)행렬, 비구조적행렬 등을 선택하여 모형을 각각 적합하였다.

Table 4.1은 AR(1)의 상관행렬을 사용한 경우의 각 설명변수의 회귀계수의 추정치와 그 때의 표준오차 그리고 회귀계수가 0인지에 대한 가설검정의 유의확률을 보여주고 있다. 질적 변수의 경우에 성별과 학년은 유의하지 않으며 부모의 교육수준인 x_9만 유의하였다. 양적 설명변수의 경우에는 월평균 사교육비 x_1과 자아존중감 x_4가 아동의 학업성취도에 긍정적 영향을 주고 그 외의 설명변수들은 유의하지 못함을 알 수 있다. 이러한 결과는 양적 설명변수들과 반응변수간의 회귀함수의 커널추정치를 보여주는 Figure 2.1에서의 결과와도 유사하다. 즉, 월평균 사교육비 x_1과 자아존중감 x_4가 증가할수록 아동 학업성취도와가 유의하게 증가하는 경향이 있어 보인다. 나머지 상관행렬들을 사용한 모형의 적합들에서도 유사한 결과가 나왔지만, 모형 선택의 한 기준인 Akaike 정보판단기준인 AIC (Akaike information criteria)의 값이 AR(1)행렬을 사용한 경우가 가장 작았기에 Table 4.1은 AR(1)행렬을 사용하였을 때의 결과들을 제시하였다. Figure 2.2에서 보여주듯이 반복 측정된 자료들의 표본상관계수들은 반복 측정의 시차가 커질수록 작아지는 경향을 보이고 있어 AR(1) 상관행렬의 형태와 유사함을 알 수 있다. 구체적 AIC 계산법은 Verbeke와 Molenberghs (2009)를 참고하기 바란다.

Table 4.1 Estimated regression parameters with their standard errors and p-values using AR(1) correlation matrix

exploratory variable		estimate	standard error	p-value
intercept		4.1940	0.6220	0.0000
gender	boy	.	.	.
	girls (x_6)	0.0267	0.1555	0.8637
grade in 2006	4^{th} grade	.	.	.
	5^{th} grade (x_7)	-0.2804	0.1914	0.1411
	6^{th} grade (x_8)	-0.1793	0.1886	0.3419
parents' educational background	low	.	.	.
	high (x_9)	0.7952	0.1694	0.0000
private education expenditure (x_1)		0.0745	0.0191	0.0001
marital and family conflict (x_2)		0.0030	0.0260	0.9079
negative parenting behavior (x_3)		-0.0460	0.0274	0.0936
self-esteem (x_4)		0.1397	0.0107	0.0000
housing area (x_5)		0.1086	0.0844	0.1984

고정효과만의 모형에서 유의한 설명변수들만 선택하여 다음과 같은 여러 모형들을 고려하였다. 즉, 아동의 학업성취도에 영향을 주는 설명변수에 대하여 각각의 고정효과와 임의효과인 임의기울기 (random slope)를 고려한 모형으로 최적의 모형을 선택하고자 한다. 모형 1은 유의한 설명변수들만 선택하

여 고정효과만을 고려한 모형이고, 모형 2는 모형 1에서 임의효과인 임의절편 (random intercept)을 추가한 모형이다. 그 이후의 모형들은 모형 2에서 유의한 양적 설명변수들인 월평균 사교육비 x_1과 자아존중감 x_4의 임의기울기를 각각 추가한 모형이다.

- 모형 1 : 유의한 설명변수들의 고정효과
- 모형 2 : 유의한 설명변수들의 고정효과 + 임의절편
- 모형 3 : 유의한 설명변수들의 고정효과 + 임의절편 + x_1의 임의기울기
- 모형 4 : 유의한 설명변수들의 고정효과 + 임의절편 + x_4의 임의기울기

네 가지 오차의 상관행렬인 단위행렬, 교환가능행렬, AR(1)행렬, 비구조적행렬을 각각의 모형에 적용하여 모형 적합을 실시하였으며 그때의 AIC들을 구하여 Table 4.2에 제시하였다. Table 4.2의 각 모형별 상관행렬별 AIC의 값을 확인해보면, 고정효과만 있는 AR(1) 상관행렬을 가정한 모형 1이 최종모형으로 타당하다는 것을 알 수 있다. 즉, 선택된 모형은 식 (3.1)에서 임의효과가 없는 다음의 식 (4.1)과 같고, 이때 오차의 상관행렬은 AR(1)이다.

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_4 x_{4ij} + \beta_9 x_{9ij} + \varepsilon_{ij}. \tag{4.1}$$

Table 4.2 AICs for each model and correlation matrix

model	1	2	3	4
identity	7559.35	7404.30	7404.38	7404.86
exchangeable	7404.30	7406.30	7406.38	7406.86
AR(1)	7383.78	7384.57	7384.02	7388.10
unstructured	7386.27	7388.28	7387.76	7389.06

Table 4.3 Estimated regression parameters of significant exploratory variables with their standard errors

significant exploratory variable		estimate	standard error
intercept		3.7844	0.4126
parents' educational background	low	.	.
	high (x_9)	0.8145	0.1677
private education expenditure (x_1)		0.0771	0.0188
self-esteem (x_4)		0.1423	0.0105

최종모형인 (4.1)의 회귀계수의 추정량과 그 때의 표준오차는 Table 4.3에 제시되어 있다. 선택된 AR(1) 상관행렬의 모상관계수 ρ의 추정치는 $\widehat{\rho}$=0.4164이다. 최종모형인 (4.1)의 모형 1과 상관행렬 AR(1)인 경우의 임의절편이 있는 모형 2의 AIC의 차이가 크지 않기에 우도비검정을 실시하여 보았다. 이 때의 검정통계량과 유의확률은 각각 1.204와 0.273으로 임의절편은 유의하지 않았다. 한편, 월평균 사교육비 x_1에 대한 임의기울기 효과가 유의한지를 검정해보기 위해 최종모형인 모형 1과 모형 3을 비교해보았다. 비교한 결과 우도비검정의 검정통계량과 유의확률은 각각 5.751과 0.124로 임의효과가 유의하지 않음을 알 수 있다. Table 4.3의 유의한 설명변수들의 추정된 회귀계수를 이용한 최종 선택된 모형 (4.1)의 추정된 회귀식은 다음과 같다.

$$\widehat{y}_{ij} = 3.7844 + 0.0771 x_{1ij} + 0.1423 x_{4ij} + 0.8145 x_{9ij}.$$

반응변수와 양적 설명변수들을 표준화한 후 최종모형 (4.1)을 적합하였을 때 월평균 사교육비 x_1과 자아존중감 x_4의 표준화된 회귀모형의 추정된 회귀계수는 각각 0.1018과 0.3022였다. 따라서 아동의 학업성취도에 자아존중감이 더 큰 영향을 미치는 것으로 파악되었다.

본 연구의 선형혼합모형의 분석 결과에서 서론에서 언급되었던 횡단자료 분석 혹은 다년도 자료의 독립적 분석을 실시한 연구들에서 아동의 학업성취도에 영향을 주는 요인들로 알려진 것들 중 부모의 경제적 지위와 관련 있는 월평균 사교육비와, 아동의 자아존중감 그리고 부모의 교육수준이 아동의 학업성취도에 영향을 준다는 것을 알 수 있었다. 반복 측정 자료이기에 각 아동의 개인적 특성이 있는지를 보고자 임의절편과 유의한 설명변수들의 임의기울기가 유의한지를 파악해보았지만 유의하지 않다는 결과를 얻게 되었다.

References

Chung, J. Y. and Jeong, Y. W. (2015). An analysis of the effects of parental involvement levels on the student achievement. *Korean Journal of Youth Studies*, **22**, 73-93.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2001). *Analysis of longitudinal data*, 2nd Ed., Oxford, New York.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, Chapman & Halls, London.

Jeon, J. Y. and Lee, K. (2014). Review and discussion of marginalized random effects models. *Journal of the Korean Data & Information Science Society*, **25**, 1263-1272.

Jo, J. and Chang, U. J. (2013). A statistical analysis of the fat mass repeated measures data using mixed model. *Journal of the Korean Data & Information Science Society*, **24**, 303-310.

Kim, B. S., Kim, D., Ha, M. and Kwon, H. (2014). Derivation of benchmark dose lower limit of lead for ADHD based on a longitudinal cohort data set. *Journal of the Korean Data & Information Science Society*, **25**, 987-998.

Kim, K. H. (2010). Effect of family income levels on academic achievement of children & adolescents : With a special focus on comparisons between child's developmental stage. *Studies on Korean Youth*, **21**, 35-65.

Kim, C. I. and Lee, K. Y. (2015). Original article : The mediating effects of ego-resilience on achievement-oriented parenting style, school adjustment and academic achievement as perceived by children. *Family and Environment Research*, **3**, 503-517.

Laird, N. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305-315.

Lim, S. and Baek, J. (2012). A credit classification method based on generalized additive models using factor scores of mixtures of common factor analyzers. *Journal of the Korean Data & Information Science Society*, **23**, 235-245.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, 1014-1022.

McCulloch, C. E., Searle, S. and Nuehaus, J. M. (2008). *Generalized linear and mixed models*, 2nd Ed., John Wiley & Sons, New Jersey.

Montgomery, D. S., Peck, E. A. and Vining, G. G. (2012). *Introduction to linear models*, 5th Ed., John Wiley & Sons, New Jersey.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-590.

Seol, K. O. and Jung, S. W. (2013). School belonging as a mediator of the relationship between individual and parental factors and academic achievement of elementary school students. *The Korean Journal of School Psychology*, **10**, 41-58.

Shim, J., Kim, Y. and Hwang, C. (2013). Generalized kernel estimating equation for panel estimation of small area unemployment rates. *Journal of the Korean Data & Information Science Society*, **24**, 1199-1210.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*, Springer Verlag, New York.

Yoo, J. Y. and Park, C. S. (2015). The effects of the parent's socioeconomic status and the private education expenditure to the academic achievement. *Journal of the Korean Data & Information Science Society*, **26**, 123-139.

Yoon, Y. H. and Choi, B. (2012). Model selection method for categorical data with non-response. *Journal of the Korean Data & Information Science Society*, **23**, 627-641.

Zhang, L. and Baek, J. (2015). The local influence of LIU type estimator in linear mixed model. *Journal of the Korean Data & Information Science Society*, **26**, 465-474.

Journal of the Korean Data & Information Science Society
2017, **28**(1), 1–10

http://dx.doi.org/10.7465/jkdi.2017.28.1.1
한국데이터정보과학회지

A longitudinal data analysis for child academic achievement with Korea welfare panel study data[†]

Naeun Lee[1] · Jib Huh[2]

[12]Department of Statistics, Duksung Women's University

Received 21 December 2016, revised 4 January 2017, accepted 5 January 2017

Abstract

Longitudinal data of Korean child academic achievement have been used to find the significant exploratory variables under the assumption of independent repeated measured data. Using the exploratory variables in previous research works, we analyze the linear mixed model incorporating the fixed and random effects for child academic achievement to detect the significant exploratory variables. Korea welfare panel study data observed three times between 2006 and 2012 by additional survey for children. The child academic achievement is evaluated by the sum of academic achievements of Korean, English and Mathematics. We also investigate the multicollinearity and the missing mechanism and select some popular correlation matrices to analyze the linear mixed model.

Keywords: Correlation matrix, fixed effect, linear mixed model, missing at random, multicollinearity, random effect.

[†] This research was supported by the Duksung Women's University Research Grants 2015.
[1] Graduate student, Department of Statistics, Duksung Women's University, Seoul 01369, Korea.
[2] Corresponding auther: Professor, Department of Statistics, Duksung Women's University, Seoul 01369, Korea. E-mail: jhuh@duksung.ac.kr