

잠재성장모형의 무조건적 모델 추정을 위한 데이터 기반 방법론

조영빈

건국대학교 국제비즈니스학부 경영학과

A Data Based Methodology for Estimating the Unconditional Model of the Latent Growth Modeling

Yeong Bin Cho

Department of Business Administration, Division of International Business, Konkuk Univ.

요 약 대표적인 종단자료 분석방법인 잠재성장모형(Latent Growth Modeling)은 무조건적 모델과 조건적 모델로 구분되는 데, 이중 무조건적 모델은 초기값과 기울기를 추정하여 적합도가 높은 모델을 추정해야 한다. 그렇지만 기존 잠재성장모형에는 종단자료의 형태가 단순선형함수 등 특정 함수가 아닐 경우 기울기를 추정하는 체계적인 방법론이 없었다. 본 연구에서는 무조건적 모델의 기울기를 추정하는데 연관규칙(Association Rule Mining)의 순차패턴(Sequential Pattern)을 사용하였다. 데이터는 한국고용정보원의 2001년~2006년에 조사한 청년 패널 데이터를 사용하였다. 제안한 방법론은 기존 단순선형 함수를 가정할 때와 비교하여 적합도가 상승하는 것을 확인할 수 있었으며, 기울기 추정 과정을 시각화할 수 있는 부수적인 장점이 있었다.

주제어 : 종단자료분석, 잠재성장모형, 무조건적 모델, 연관규칙, 순차패턴

Abstract The Latent Growth Modeling(LGM) is known as the arising analysis method of longitudinal data and it could be classified into unconditional model and conditional model. Unconditional model requires estimated value of intercept and slope to complete a model of fitness. However, the existing LGM is in absence of a structured methodology to estimate slope when longitudinal data is neither simple linear function nor the pre-defined function. This study used Sequential Pattern of Association Rule Mining to calculate slope of unconditional model. The applied dataset is 'the Youth Panel 2001-2006' from Korea Employment Information Service. The proposed methodology was able to identify increasing fitness of the model comparing to the existing simple linear function and visualizing process of slope estimation.

Key Words : Longitudinal Data Analysis, Latent Growth Modeling, Unconditional Model, Association Rule, Sequential Pattern

1. 서론

빅 데이터 시대가 되면서 수많은 데이터가 축적되고 확대되고 있다. 빅 데이터의 주류를 이루는 비정형 데이터뿐만 아니라 정형화된 데이터도 지속적으로 확대 저장

되고 있다. 그 중 정기적인 조사를 거쳐 수집되는 종단 자료도 많아지고 있는데, 우리나라에서는 교육, 의료, 노동 등의 분야에서 청소년 패널자료, 교육종단자료, 노동 패널자료, 아동패널자료 등 종단 자료가 구축되어 공개되고 있다. 해외에서는 교육학 분야를 중심으로 다양한

*Corresponding Author : Yeong Bin Cho(ybcho111@kku.ac.kr)

Received April 4, 2018

Accepted June 20, 2018

Revised May 16, 2018

Published June 28, 2018

항목을 가진 종단자료가 수집되고 있다(Early Childhood Longitudinal Study, National Education Longitudinal Study, Longitudinal Study of American Youth 등)[1]. 경영학 분야에서도 재무 및 마케팅 분야를 중심으로 종단 연구가 이루어지고 있다[2-6, 19, 20].

여러 종단자료 분석방법 중 최근 많은 관심을 받고 있는 방법은 잠재성장모형(Latent Growth Modeling, LGM)이다. 잠재성장모형은 구조방정식모형과 유사한 분석 틀로 구성되어 있어, 잠재변인(Latent variable)과 관측변인(Observed variable)을 구분하여 분석한다. 잠재변인은 기울기와 초기값에 해당하고 관측변인은 매회 측정된 종단자료를 사용한다. 잠재성장모형을 사용하는 절차는 먼저 변수들의 종단적인 변화추이를 분석한다. 다음으로는 ① 무조건적 모델(Unconditional Latent Curve Model)의 초기값과 기울기를 추정하여 적합도가 높은 모델을 찾고 ② 독립변인을 포함한 조건적 모델(Conditional Latent Curve Model)을 만들어 시간의 변화를 반영한 변수를 반영하여 종단 인과관계 검증을 하게 된다[7]. 이러한 모델 구축은 매우 어려운데, 왜냐하면 정형화된 방법론이 없고 시행착오적(trial and error)방법에 의해서 찾기 때문이다[8]. 다시 말해 무조건적 모델의 모수를 추정하고 나서 조건적 모델을 구성해야 종단관계의 인과 검증이 가능하기 때문이다.

본 연구의 범위는 잠재성장모형의 모형 적합도를 향상시키기 위하여 무조건적 모델의 기울기를 추정하는 방법론을 제시한다. 기존 방법론은 단순 선형함수, 2차 혹은 지수함수 등을 가정하여 기울기를 추정해왔다. 그렇지만 종단자료의 형태가 가정한 특정 함수와 다를 경우 모형 적합도를 향상시킬 수 있는 방법론이 없는 상태이다[7, 9]. 본 연구에서는 무조건적 모델의 기울기 추정에 대표적인 데이터 마이닝 기법인 연관 규칙(Association Rule)의 순차패턴(Sequential Pattern)을 이용하였다. 일정 수준이상의 지지도(Support)와 신뢰도(Confidence) 이상의 시계열 패턴을 찾아서 이를 순차별로 정리하여 무조건적 모델의 기울기 추정에 사용하였다.

제시한 방법론을 적용한 데이터는 한국고용정보원의 2001년~2006년에 조사한 청년 패널 데이터로서, 청년 패널 조사는 15~29세의 청년층을 대상으로 직업선택 및 노동시장 이동 등을 조사한 자료이다. 공개된 종단 데이터를 사용하여 제시한 방법론을 실험해본 결과, 기존 단 순선형함수를 가정할 때와 비교하여 모형적합도가 상승

하는 것을 확인할 수 있었으며, 기울기 추정 과정을 시각화할 수 있는 부수적인 장점이 있었다. 잠재성장모형과 데이터마이닝 기법을 결합할 수 있는 계기를 마련하였다.

2. 관련 연구 및 방법론

2.1 종단연구(Longitudinal Study)

종단자료(Longitudinal Data)란 전체 연구기간 동안 여러 번 정보를 수집하는 경우 발생하는 정보를 말한다. 횡단자료(Cross-sectional data)는 전체 연구기간 동안 특정 시점에 정보가 수집되는 경우이다. 전체 연구기간 동안 연속적인 횡단연구가 이루어 질 경우 수집되는 자료는 종단자료와 같다. 그렇지만 자료 구성면에서 종단 자료는 적어도 3번의 자료가 시간의 변화에 따라 측정되어야 한다[11]. 2회의 자료만으로는 종단자료의 분석이 불가능하다. 2회 측정만으로는 측정오차와 변화량을 구분할 수 없기 때문이다.

종단연구는 종단자료를 대상으로 하는 연구방법이다. 종단연구의 가장 커다란 장점은 인과관계 증명이 용이하다는 것이다. 인과관계 증명을 위해서는 3가지 조건을 만족해야 함은 잘 알려져 있다. 첫 번째는 공변량(Covariation)으로, 관심이 있는 두 변수 간 유의한 연관성(Association)이 있어야 한다. 두 번째는 진실한 관계(Non-spuriousness)로, 두 변수간의 연관성은 다른 변수의 영향에 기인하지 않는다. 세 번째는 사건의 순서(Temporal order of events)로, 원인변수는 영향변수 보다 반드시 앞서 일어나야 한다는 것이다. 이중 종단연구는 세 번째 인과 조건을 명확하게 증명할 수 있다. 반면 종단연구는 자료의 양이 매우 많고, 수집비용이 많이 든다는 단점이 있다[12].

대부분의 종단 연구는 시간흐름에 따른 종속변수의 변화를 추정하고 나서 종속변수에 영향을 주는 독립변수도 추정한다. 시간흐름에 의한 변수들의 변화량과 종단 연구의 대상이 되는 개인 내 변화(Within-individual change)를 추정한다[13]. 연구 대상에 따라 동일한 대상을 반복적으로 측정하는 패널(Panel)조사와 동일 조건의 대상물을 반복 측정하는 코호트(Cohort) 조사, 두 개 이상의 횡단분석이 두 개 이상의 시간에서 실시되는 경향 분석(Trend Analysis), 실험군에 처치를 가하기 전과 가한 후를 비교하는 반복측정 분산분석(Repeated measures

ANOVA) 등이 있다[12].

2.2 잠재성장모형(Latent Growth Model)

잠재성장모형은 1987년 McArdle와 Esptein[14]의 연구자들에 의해서 처음으로 제안된 종단자료 분석방법이다. 또한 잠재성장모형은 구조방정식모형(Structural Equation Modeling: SEM)의 특별한 유형으로 설명될 수도 있다[1, 15]. 구조방정식모형 중 잠재변인(Latent variable)과 관측변인(Observed variable)은 중요한 개념 중 한가지이다[16]. 잠재변인과 관측변인은 잠재성장모형에서도 동일하게 적용된다. 잠재성장모형에서 잠재변인은 기울기(Slope)와 초기값(Initial Status)에 해당되고, 관측변인은 매회에 측정된 종단자료들로 정의된다.

일반적인 잠재성장모형을 방정식으로 표현하면 수식 (1)과 같이 표현할 수 있다.

$$y_{it} = \beta_{0i} + \beta_{1i}x_t + \varepsilon_{it} \quad (1)$$

수식 (1)에서 y_{it} 는 i 번째 패널의 종단자료 중 t 주기에 측정된 종단자료 값을 의미하며, β_{0i} 는 패널 i 의 초기값을 의미한다. β_{1i} 는 i 번째 패널의 기울기 값을 나타내며, x_t 는 t 측정주기의 기울기 계수값을 나타낸다[17]. 수식 (1)은 2개의 하위 계층으로 분리되어 개별효과를 추정하는데 사용된다.

수식 (1)을 두 개의 공식으로 분리하면 다음과 같다.

$$\beta_{0i} = \beta_0 + \zeta_{0i} \quad (2)$$

$$\beta_{1i} = \beta_1 + \zeta_{1i} \quad (3)$$

수식 (2)에서 β_0 는 전체 평균의 초기값을 의미하고 ζ_{0i} 는 전체평균과 i 번째 패널 간의 차이를 나타낸다. 마찬가지로 수식 (3)은 수식 (1)의 개별 패널의 기울기 값을 전체 평균의 기울기(β_1)와 전체평균과 개인의 차이(ζ_{1i}) 점수로 구분하고 있다. 수식의 의미는 전체 평균의 기울기와 개별 패널의 기울기 간 차이는 존재할 수 있으며, 이러한 차이는 개별 패널의 특수성으로 추출될 수 있다. 따라서 전체 평균의 기울기 뿐만 아니라, 집단에 소속된 개별 패널들의 기울기의 분산이 통계적으로 유의한지를 밝혀서 전체적인 변화와 개별 패널의 시간 흐름에 따른 변화를 모두 추정할 수 있다[17].

잠재성장모형의 분석은 무조건적 모델과 조건적 모델

의 분석으로 구분한다. 무조건적 모델은 독립변인을 포함하지 않은 모형을 의미한다. 무조건적 모델을 분석한 후에 기울기와 초기값의 분산이 통계적으로 유의할 경우 추가적인 조건적 모델에 대한 추정을 실시하게 된다[9]. 무조건적 모델의 기울기와 초기값을 추정하는데 있어서 뚜렷한 방법론이 없다. 가장 많이 사용되는 방법은 종속변수의 대략적인 추세를 추정한 후 기울기를 찾는 방법이다. 그런데 이 방법은 종속변수의 추세선이 선형함수에 부합할 경우에는 적절하지만, 비선형 추세선일 경우 휴리스틱한 방법을 사용하게 된다. 적절한 기울기 값을 가정하여 무조건적 모델의 기울기를 찾아내게 된다. 무조건적 모델이 완성된 이후에는 조건적 모델을 추정하게 된다. 조건적 모델을 만드는데 가장 중요한 과정은 적합도가 높은 모델을 찾는 것인데, 이는 시간 흐름에 따라 반응의 변화를 추정하고 패널 내 변화 중 의미 있는 변화를 추정하는 과정과 같다. 다시 말해 시간의 흐름에 따라 변화하는 종단적 궤적(Longitudinal Trajectory)을 찾아내는 일이다[9].

2.3 연관(순차패턴) 마이닝

(Association[sequential pattern] Mining)

연관 규칙(Association Rules) 기법은 사건들 사이의 상호 연관성 추정에 사용되는 기법이다. 연관규칙 기법을 이용하여 추출되는 결과는 'X사건이 일어나면 Y사건도 일어 난다'는 의미이고 다음과 같은 규칙(rule)형태로 표현한다.

$$\text{If } X, \text{ Then } Y (X \rightarrow Y) \quad (4)$$

연관규칙은 장바구니에 담긴 아이템의 상호 연관성을 찾기 위하여 개발된 방법이며, 연관규칙을 찾기 위하여 주로 사용되는 척도는 지지도(Support), 신뢰도(Confidence)이다. 지지도(Support)는 생성된 연관규칙이 전체 항목에서 차지하는 비율을 의미한다. 즉, 전체 거래 중 그 연관규칙을 지지하는 거래의 비율을 의미한다. 신뢰도(Confidence)는 X를 포함하는 거래 중에서 Y가 포함된 거래의 정도를 의미하며 연관규칙의 강도를 의미한다. 규칙의 왼쪽인 조건부분을 만족하는 거래 중 규칙의 오른쪽인 결론 부분까지 만족하는 비율을 말한다. 신뢰도가 규칙의 강도를 나타낸다면, 지지도는 이 규칙이 전체 데이터베이스에서 가지는 통계적인 중요성을 나타낸다고 할 수 있다. 연관규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발되었고, 가장 많이 사용되고 있는 것은

Apriori 알고리즘이다[10]. 이러한 동시 출현빈도를 찾는 과정은 데이터베이스에서 빈번항목을 추출하는 과정으로서 이 알고리즘은 두가지 단계로 구성된다. 첫 번째 단계는 최소 지지도 설정 값에 따라 빈도수가 높은 항목의 집합들을 찾아내고, 이들 집합들 중에서 주어진 신뢰도를 만족하는 연관규칙을 찾아낸다[18].

순차패턴 마이닝(Sequential Pattern Mining)은 연관 규칙에 시간의 흐름에 따른 항목들의 상호 연관성을 탐색하는 것이다. 순차패턴(Sequential Pattern)은 연관 규칙에 시간 개념을 첨가하여 시계열 패턴들의 상호 연관성을 탐색하는 기법이다. 연관 규칙과 유사하지만 거래 데이터베이스로부터 고객이 선호하는 상품 항목(데이터 시퀀스)들의 순차 패턴을 추출하여 순차(Sequence) 데이터베이스를 생성한다. 순차 패턴 기법은 거래 발생 순서에 따라 각 거래에서 사용되는 상품 항목들의 연관성을 탐색하여 거래 간에 영향을 미치는 선행 항목을 검색한다. 즉 고객의 상품 구매 패턴을 분석하여 향후 구매 가능한 상품 예측에서 사용된다[18].

순차패턴 발견에서의 연관 규칙 $X \rightarrow Y$ 는 “상품 X가 구매되면 일정기간이 경과한 다음에 상품 Y가 구매된다.”라고 해석한다. 순차적 패턴발견은 구매 순서가 고려되어 상품간의 연관성이 측정되고 연관성의 정도에 따라 유용한 연관규칙을 찾는 기법이다. 연관 규칙과 순차패턴이 다른 점은 연관 규칙은 $X \rightarrow Y$, $Y \rightarrow X$ 가 성립하지만, 순차패턴은 $X \rightarrow Y$ 만 성립한다는 것이다. 다시 말해 연관규칙은 X, Y 중 어느 것이 먼저 일어나도 관계 없지만 순차패턴은 반드시 X가 먼저 발생하는 것을 말한다.

3. 잠재성장모형 분석방법

3.1 적용 종단자료 및 기술통계

본 연구에서 사용된 종단자료는 한국고용정보원의 2001년~2006년에 조사한 청년 패널 데이터로서, 청년 패널 조사는 15~29세의 청년층을 대상으로 직업선택 및 노동시장 이동 등을 조사한 자료이다. 분석목적에 맞게 발췌하여 dataset을 구성하였다. 최근 종단자료 분석을 할 수 있는 다양한 통계 프로그램(예, AMOS, STATA, LISREL)이 소개되고 있으며, 이러한 프로그램은 시각적인 도형으로 모형을 설계하고 결과를 쉽고 정확하게 해석할 수 있도록 고안되었다. 본 연구에서는 SPSS

AMOS 24와 STATA프로그램을 사용하였다.

샘플 수는 76명이며, 6년 동안 종단조사를 실시하였기 때문에 총 샘플은 456개이다. 2001년부터 2006년까지 6년간 임금과 임금에 영향을 미치는 11개의 독립변수가 수집되었다. Fig. 1에서 제시하는 바와 같이 결측치는 사전에 제거하여 종단분석을 위한 전처리를 실시하였다.

pid: 1104, 2804, ..., 173303	n = 76
year: 2001, 2002, ..., 2006	T = 6
Delta(year) = 1 unit	
Span(year) = 6 periods	
(pid*year uniquely identifies each observation)	

Fig. 1. general characteristics of the longitudinal data

3.2 무조건적 모델 분석

SPSS AMOS에서도 잠재성장모형의 분석은 무조건적 모델과 조건적 모델 분석으로 구분한다. 무조건적 모델이란 독립변수가 포함되지 않은 종속 변수만의 변화모형을 의미한다. 무조건적 모델을 분석한 후에 기울기와 초기 값의 분산이 통계적으로 유의할 경우 조건적 모델의 분석을 실시하게 된다. 기울기와 초기 값의 분산이 유의하다는 의미는 개별 패널들의 변화가 동일하지 않다는 것을 의미한다. 따라서 패널 간 변화 유형의 차이를 설명하는 조건적 모델 분석을 실시하게 된다.

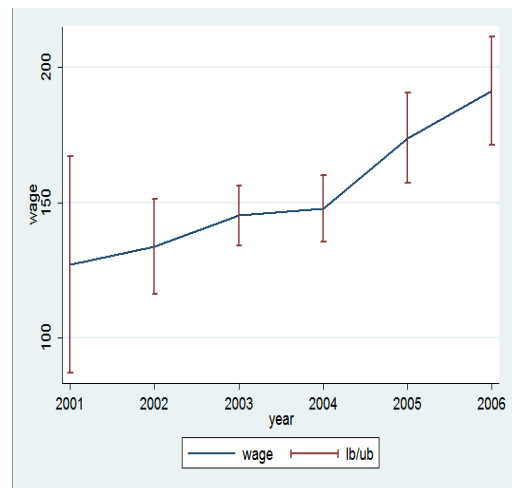


Fig. 2. longitudinal trajectory of 'wage' variable

Figure. 2.는 종속변수인 임금 수준의 종단 추이를 나타낸다. 먼저 기울기는 2001년부터 2003년까지 완전한

상승세를 보이다가 2004년에는 2003년과 비슷한 수준을 유지하였고, 2005년, 2006년에는 상승폭이 커지는 것을 발견할 수 있다. 수직선은 95% 신뢰구간의 상한치와 하한치를 나타낸다. 2001년의 패널별 평균임금의 변동 폭이 가장 크고 2003년이 가장 적은 것을 알 수 있다. 이는 연도별 표준편차가 서로 상이하다는 것을 의미한다. Table 1에는 연도별 임금 평균과 95% 신뢰구간의 상한치, 하한치가 제시되어 있다.

Table 1. Arithmetic mean of wage with bars based on 95% CI by year.

	2001	2002	2003	2004	2005	2006
mean	127.2	133.8	145.2	147.8	173.9	191.2
lower	87.2	116.2	134.2	135.5	157.1	171.2
upper	167.2	151.3	156.2	160.1	190.6	211.2
gap	80.0	35.1	22.0	24.6	33.5	40.0

Fig. 3에는 SPSS AMOS 잠재성장모형 분석 화면을 나타낸다. 무조건적 모델을 분석하기 위해서는 모든 변인들을 잠재성장모형에 투입한 후 관측변인과 잠재변인 간의 계수를 설정해주어야 한다. SPSS AMOS에서는 화살표로 표현된 경로계수에 적절한 값을 입력하도록 되어 있다. 본 연구의 대상인 데이터 셋에서 6개의 종단자료와 초기값(ICEPT)은 시간의 변화와 관계없이 일정한 값을 가지고 있다. 즉 처음 측정된 값은 그 이후에도 변화가 없기 때문에 6개의 관측변인과 초기값의 경로계수는 1로 지정된다.

잠재성장모형 분석에서 기울기(SLOPE) 경로계수 설정은 특별한 방법이 지정되어 있지 않다. 일반적으로는 선형함수, 2차 함수, 지수함수 등 함수를 가정하여 계수를 입력하도록 한다. 선형함수로 가정할 경우 6기 동안의 경로계수로 0, 1, 2, 3, 4, 5, 6을 입력하면 된다. 그렇지만 Fig. 2와 같이 데이터 셋의 기울기는 전반적으로 증가세를 보이고 있지만, 전반기의 기울기와 후반기의 기울기가 서로 다른 형태를 띠고 있다. 또한 매 기의 표준편차도 서로 상이 하여 단순한 선형함수로 가정하기는 어렵다. 따라서 종단자료의 종속변수의 형태가 일반적인 선형함수가 아닌 경우에 이를 해결할 다른 방법이 필요하다.

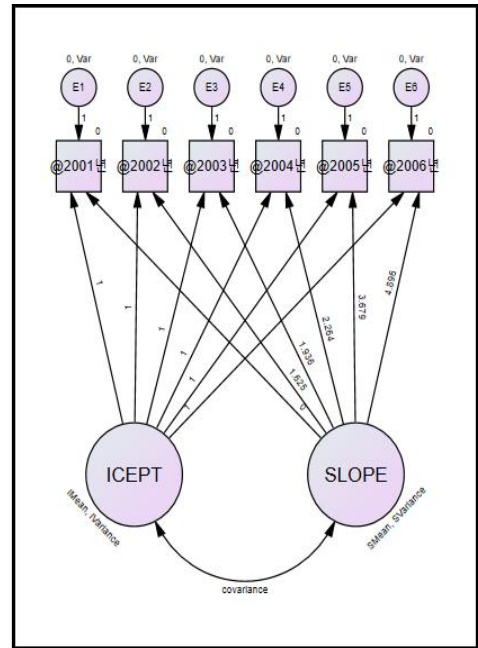


Fig. 3. Unconditional model of Latent Growth Model in SPSS AMOS

4. 연관규칙 기법의 적용

4.1 순차패턴 추출

전술한 바와 같이 연관규칙 마이닝 기법은 $X \rightarrow Y$ 형태의 규칙을 생성한다. 이를 종단 자료에 적용하면 $(X_1, X_2, \dots, X_n) \rightarrow Y$ 의 형태의 규칙을 생성하게 된다. 이는 종단자료에 대한 종단적 궤적을 추출하는데 사용될 수 있다. 그렇지만 통계적인 타당성이 떨어지는 종단적 궤적은 무조건적 모델의 기울기 추정에 부적절할 수 있다. 왜냐하면 그러한 궤적은 특정 패널의 궤적만을 반영하기 때문이다. 연관규칙(association rule)은 자동화된 필터링 기능을 갖고 있으면서 신뢰성이 높은 규칙을 도출할 수 있기 때문에 이러한 상황에서 사용하기에 적절한 기법이라 할 수 있다. SPSS Modeler의 연관규칙 모듈을 사용하여 연관 규칙을 도출하였다.

T시점에 1 순차이면서 최소 지지도와 신뢰도이상의 연관규칙을 R_j 라 하면 R_j 는 다음과 같이 나타낼 수 있다.

$$R_j : r_{j, T-l+1}, \dots, r_{j, T-1} \rightarrow r_{j, T} \quad (\text{support}_j, \text{confidence}_j) \quad (5)$$

예를 들어 2003년에 3순차이면서 지지도가 10%이고 신뢰도가 50%인 규칙은 $r_{2001}, r_{2002} \rightarrow r_{2003}$ (10%, 50%)와 같이 나타낼 수 있다.

연관규칙을 사용하려면 실수형 척도를 서열형이나 범주형 척도로 전환하여야 한다. 본 연구 대상인 데이터 셋의 종속변수는 실수형 척도이므로 이를 범주형 척도로 전환하였다. SPSS Modeler에서 제공한 기능을 사용하여 wage를 분위 수(fractile)로 전환하였으며, 각각 4분위, 5분위, 10분위로 전환하였다. 따라서 wage는 각각 4, 5, 10개의 범주를 가진 3개의 변수로 치환되었다. 치환된 변수를 이용하여 연도별 순차패턴을 추출하였다.

Fig. 4 - Fig. 8은 산출된 순차패턴을 나타낸다. 지지도와 신뢰도는 연구자의 필요에 의해서 결정할 수 있으나[10], 본 연구에서는 지지도는 10%이상, 신뢰도는 50% 이상을 사용하였다. 예를 들어 Fig. 4의 첫 번째 순차패턴은 2001년에 1분위였던 패널이 2002년에도 1분위인 규칙이 산출되었으며, 이 규칙의 지지도는 21.05%, 신뢰도는 56.25%임을 나타낸다. 또한 2001년과 2002년 사이에 산출된 규칙은 1분위 간 1개, 4분위 간 1개로 총 2개였다. Fig. 5의 3순차 패턴규칙 중 첫 번째 규칙은 2001년에 1분위, 2002년에 1분위인 패널이 2003년에도 1분위일 가능성이 높다는 것을 의미하고, 그 가능성 정도는 지지도 11.84%, 신뢰도 88.89%임을 보여준다.

4.2 기율기 추정

산출된 순차패턴은 4분위에서 2순차, 3순차, 4순차패턴으로 3개, 5분위에서 2순차, 3순차로 2개, 10분위에서 2순차 등 총 6개이다. 산출된 순차패턴규칙은 연도별 해당 분위 변화량에 대한 가중평균으로 사용할 수 있다.

전술한바와 같이 4분위 수의 경우 2001년 1분위에서 2002년 1분위로 이동하는 규칙 일정 지지도와 신뢰도를 넘어 산출되었다. 이는 2001년 1분위 상태에서 2002년 1분위 상태로 이동하는 패널의 숫자가 뚜렷한 패턴이 발견되었으므로, 가중평균으로 사용할 수 있음을 나타낸다.

	2001	2002	2003	2004	2005	2006
1 frac.		(21.05, 56.25)	(23.68, 55.58)	(21.05, 68.75)	(21.05, 81.25)	
2 frac.		(18.42, 71.43)		(21.05, 68.75)		
3 frac.					(22.37, 58.82)	
4 frac.	(28.95, 77.27)	(27.63, 90.48)	(27.63, 71.43)	(26.31, 70.00)	(28.95, 77.27)	

Fig. 4. 4-fractile 2-sequences association (sequential pattern) rules

	2001	2002	2003	2004	2005	2006
1 frac.		(11.84, 88.89)		(10.53, 87.50)	(14.47, 90.91)	
2 frac.		(13.16, 50.00)	(13.16, 50.00)			
3 frac.				(10.53, 75.00)		
4 frac.			(19.74, 73.33)	(25.00, 78.95)		

Fig. 5. 4-fractile 3-sequences and 4-sequences association (sequential pattern) rules

	2001	2002	2003	2004	2005	2006
1 frac.	(18.42, 57.14)		(11.84, 77.78)	(13.16, 50.00)	(13.16, 90.00)	
2 frac.					(23.68, 50.00)	
3 frac.						
4 frac.	(19.74, 53.33)	(21.05, 81.25)	(23.68, 66.67)	(21.05, 62.50)		
5 frac.	(21.05, 62.50)				(21.05, 81.25)	

Fig. 6. 5-fractile 2-sequences association (sequential pattern) rules

	2001	2002	2003	2004	2005	2006
1 frac.						
2 frac.						
3 frac.						
4 frac.			(10.52, 50.00)			
5 frac.			(15.79, 75.00)			

Fig. 7. 5-fractile 3-sequences association (sequential pattern) rules

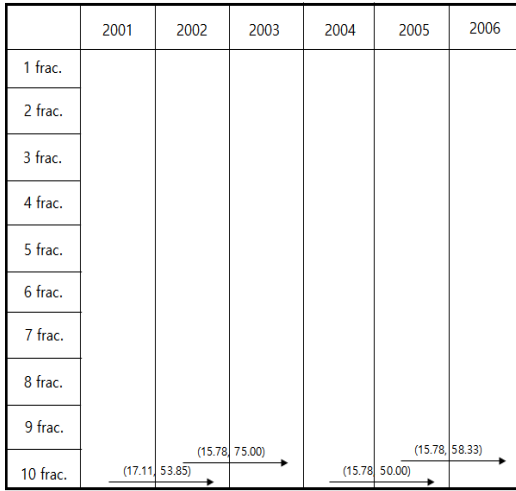


Fig. 8. 10-fractile 2-sequences association (sequential pattern) rules

식(6)과 같이 본 연구에서는 다음과 같은 가중평균을 이용하여 순차별 기울기를 산출하였다.

$$\sum_{i=1}^T \Delta_l \times R_j(S_j \times C_j) \quad (6)$$

여기서 T : 분위 수, l : 순차, Δ_l : 상위 연도 분위 평균 - 하위 연도 분위 평균, R_j : j 번째 순차규칙, S_j : j 번째 순차규칙의 지지도, C_j : j 번째 순차규칙의 신뢰도

이렇게 기울기에 순차패턴을 적용하면 단순선형모형이 아닌 비선형모형을 산출하게 된다. 다시 말해 연도별 기울기가 달라지면서 종단자료의 종단적 궤적에 더 근접하는 결과가 된다.

4.3 분석결과

무조건적 모델의 초기치와 기울기를 입력한 후 모형 적합성을 평가하기 위하여 실제 종단자료와의 적합도 정도를 평가한다. 즉 가설된 연구모형이 종단자료의 특성을 잘 반영하고 있는지를 평가해야한다. 적합도 평가방법은 CFI(Comparative Fit Index), NFI(Normed Fit Index), IFI (Incremental Fit Index), TLI (Tucker-Lewis Index)와 같은 절대적 적합도 지수와 상대적 적합도 지수인 χ^2 검사 값을 사용하면 된다. 절대적 적합도 지수의 경우 0에서 1까지의 값을 가질 수 있고 값이 클수록 적합도가 좋은 것으로 알려져 있다. χ^2 적합도 지수는

적으면 적을수록 적합도가 우수하다고 평가할 수 있지만, 절대적인 기준 값을 제시할 수 없는 문제점이 있다[9].

6개 순차패턴을 이용하여 추정한 기울기와 단순선형 모형의 적합도 지수가 Table 2에 제시되어 있다. 분석 결과 6개 순차패턴 모두 단순선형모형보다 적합도가 좋은 것으로 나타났다. 가장 적합도가 높은 순차패턴은 4분위 2순차패턴으로 나타났다. 4분위 2순차 패턴 모델에 대한 기울기와 초기값의 평균과 분산에 대한 추정치와 표준오차는 Table 3에 제시하였다.

Table 2. model fitness of simple linear function and derived sequential patterns

		model fit					BM
fit index		χ^2	CFI	NFI	IFI	TLI	
4-frac.	simple linear	172.4	.374	.329	.358	.553	BM
	2seq.	162.0	.417	.369	.402	.584	**
	3seq.	163.9	.409	.362	.394	.578	
	4seq.	167.1	.397	.350	.381	.569	
5-frac.	2seq.	165.5	.403	.356	.388	.573	
	3seq.	165.2	.404	.357	.389	.574	
10-frac.	2seq.	170.2	.383	.337	.367	.559	

산출된 초기값과 기울기 값의 분산은 모두 통계적으로 유의한 것으로 나타났다.

Table 3. result of initial status and slope

		4-fractile 2-sequence nonlinear model			
		mean		variance	
		estimate	SD	estimate	SD
wage	icpt	111.4*	5.2	1056.3*	334.6
	slope	14.2*	1.9	196.9*	46.9

*: $p < 0.001$

5. 결론 및 시사점

잠재성장모형은 기존 분석방법에 비해 오랜 학습이 필요하다. 종단자료 개념, 통계적인 기초지식 뿐만 구조 방정식과 위계적 선형모형 등 복잡한 분석기법도 학습해야하기 때문이다. 그렇지만 빅데이터가 강조되면서 축적되는 데이터의 양과 질이 급격히 증가하고 있는 최근의 상황에 비추어 볼 때 종단자료에 대한 분석이 더욱 활성화될 것 이다.

본 연구의 의의는 잠재성장모형 중 무조건적 모델의 기울기 추정문제에 새로운 방법론을 제시하였고, 제시한 방법론이 종단자료와 무조건적 모델의 적합성을 제고했다는 데 있다. 이지선, 김세영[1]의 연구처럼 수학적 방법으로 잠재성장모형의 방법론적 효율성을 제고하기 보다는 휴리스틱한 방법을 사용했다는데 의의가 있다. 또한 기존 방법론[7, 9]은 종속변수의 종단적 궤적을 분석한 후 연구자의 직관에 따라 기울기를 정한 반면 본 연구에서는 체계적인 방법론을 제시하였다. 두 번째는 종단자료분석을 위한 잠재성장모형에 기존의 데이터 마이닝 기법을 적용하여 성능을 향상했다는 데 있다. 기존 연구[1, 7, 9, 11]에서는 통계적인 방법을 이용하여 성능향상을 시도하였지만, 본 연구에서는 연관규칙이라는 데이터마이닝 기법을 적용하여 적용기법의 범위를 확장하였다. 세 번째로 본 연구의 방법론 중 연관규칙의 순차패턴은 데이터 시각화(Data Visualization) 도구로 사용될 수 있을 것이다. 이를 사용할 경우 SPSS 통계패키지의 성능 향상에도 도움이 될 것으로 예상된다.

본 연구의 한계는 연관규칙의 지지도와 신뢰도를 선정하는데 있어서 경험적인 수치를 사용하였다는 데 있다. 또한 본 연구에서 사용한 데이터는 단순적인 비선형 모델이 아니라 2차 함수나 지수함수 형태가 아닌지 검토할 필요가 있다. 두 번째로는 한국고용정보원의 한정된 데이터를 사용하여 시행한 탐색적 연구(Exploratory Study)이기 때문에 다른 다양한 데이터를 사용하여 본 연구에서의 방법론을 시험할 필요가 있다.

추가적인 연구로는 방법론으로만 제한하지 말고 이론적인 배경에 따른 행태적인 연구를 같이 하면서, 기술적인 성능 향상과 함께 이론 검증과 새로운 사실발견도 포함하는 종합적인 연구가 필요할 것이다.

REFERENCES

- [1] J. S. Lee & S. Y. Kim. (2017). An Exploration of Nonlinear Latent Growth Model Using Exponential Function: As an Alternative to Quadratic LGM, *J. of Educational Evaluation*, 30(4), 791-816.
- [2] B. W. Jun, Y. W. Joo & B. H. Lim. (2011). A Study of Demographic Characteristics and Products Buying Behavior of Online Consumer Survey Participants Compared to Face-to-Face Survey, *Korean Coporation Management Review*, 21(3), 93-108.
- [3] E. J. Lee & C. H. Cho. (2013). A Longitudinal Study on the Effects of Franchise's Factors and Perform ance - Disclosure Agreement, *Korean J. of Business Administration*, 26(8), 2185-2209.
- [4] E. J. Lee & C. H. Cho. (2014). A Logitudinal Study on the KS-SQI for Improving Service Quality, *Korean J. of Business Administration*, 27(4), 561-574.
- [5] H. J. Lim & J. S. Cho. (2012). The Effect of Ownership Concentration on Firm Performance : Static and Dynamic Panel Data Analysis, *Korean J. of Business Administration*, 25(8), 3265-3291.
- [6] Y. B. Cho, S. K. Lee & K. H. Ro. (2015). A Methodology for Analyzing the Longitudinal Data using SOM Technique, *Korean J. of Business Administration*, 28(1), 93-102.
- [7] K. S. Kim. (2009). *AMOS and LISREL*, Han Academy.
- [8] S. W. Menard. (2002). *Longitudinal research (2nd. ed.)*. London: Sage Publications Inc.
- [9] S. S. Yeo & S. H. Park. (2012). An Appliation of Latent Growth Modeling: Use of Curriculum-Based Measurement as longitudinal Data, *Asian J. of Education*, 13(4), 247-273.
- [10] R Agrawal, T. Imielinski & A. Swami. (1993). Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.
- [11] J. D. Singer & J. B. Willett. (2006). *Longitudinal data analysis: present status; future prospects*, Presentation at the 45th Congress of the German Pycological Association, Nurnberg, Germany, 17-21.
- [12] Toon Taris. (1999). *A Primer in Longitudinal Data Analysis*, SAGE Publications Inc.
- [13] G. M. Fitzmaurice, N. M. Laird & J. H. Ware. (2012). *Applied Longitudinal Analysis*, 2nded. JohnWiley&Sons, Hoboken, New Jersey.
- [14] K. L. McArdle & D. B. Epstein. (1987). Latent Growth curves within development structural equation models, *Child Development*, 58, 110-133.
- [15] B. M. Byrne. (2010). *Structural Equation modeling with AMOS: Basic concepts, applications and programming*. Mahwah, N.J:Lawrence Erlbaum Associates.
- [16] R. B. Kline. (2004). *Principles and practice of structural equation modeling*. New York: Guilford.
- [17] K. A. Bollen & P. J. Curran. (2006). *Latent curve models a structural equation perspective*. Hoboken, NJ: Wiley-Interscience.
- [18] B. W. Jin, Y. S. Cho & K. H. Ryu. (2010). Personalized e-Commerce Recommendation System using RFM

- method and Association Rules, *J. of the Korea Society of Computer and Information*, 15(12), 227-235.
- [19] J. C. Kim, H. I. Jung, H. Yoo & K. Y. Chung. (2018). Sequence Mining based Manufacturing Process using Decision Model in Cognitive Factory, *Journal of the Korea Convergence Society*, 9(3), 53-59.
- [20] Y. J. Shin & M. S. Yim. (2012). A Study of the Relationship Analysis between Mobile Application by Using An Association Rules, *Journal of the Korea Convergence Society*, 3(2), 19-25.

조 영 빈(Cho, Yeong Bin)

[정회원]



- 1985년 2월 : 고려대학교 산업공학과(공학사)
- 1988년 2월 : 한국과학기술원 산업공학과(공학석사)
- 2005년 2월 : 한국과학기술원 경영대학 경영공학(경영정보학 박사)
- 2006년 3월 ~ 현재 : 건국대학교 글로벌캠퍼스 국제비즈니스학부 경영학과 교수
- 관심분야 : CRM, 데이터마이닝, 온라인고객
- E-Mail : ybcho111@kku.ac.kr