

Group lasso를 통한 중학생의 삶의 만족도에 영향을 미치는 변수 탐색

유진은* · 노민정**

초 록

한국청소년정책연구원의 KCYPs 패널 자료가 수백 개의 변수들을 제공함에도, 선행 연구들은 그 중 십여 개의 변수 또는 요인만을 선정하여 청소년의 삶의 만족도를 모형화해왔다. 그러나 복잡다단한 인간을 대상으로 하는 사회과학 연구는 기존 이론에 근거한 몇 가지 변수 또는 요인만을 이용하는 데서 벗어나, 다른 새로운 변수들을 파악함으로써 기존 이론을 정비할 필요가 있다. 특히 KCYPs가 제공하는 변수들은 그 이론적·실제적 타당성을 다수의 전문가들이 연구 설계에서부터 고려하여 수집된 변수들이다. 따라서 본 연구는 KCYPs 6차년도 초4 패널 자료 전체를 이용하여 중학생의 삶의 만족도에 영향을 미치는 변수를 탐색하고자 별점 회귀모형 기법인 group lasso를 적용하였다. 별점 회귀모형은 비별점 회귀모형이 투입변수가 증가함에 따라 모형안정성이 떨어지는 문제에 대하여 해결책을 제시한다. 또한 본 연구에서 group lasso를 적용함으로써 모형에서 연속형 변수뿐만 아니라 범주형 변수를 함께 다룰 수 있었다. 그 결과 338개 변수 중 15 개 변수가 선택되었으며, 이 때 모형의 예측 정확도는 78.25%였다. 선택된 변수 중 자아인식, 정서문제, 양육방식, 전체성적 만족도, 지역사회인식, 가정경제수준 평가 관련 문항은 선행연구에서 이용된 바 있으며, 그 회귀계수 방향 또한 선행연구 결과와 전반적으로 일치하였다. 선행연구에서 다루어지지 않았으나 별점 회귀모형에서 선택된 변수로 공부시간, 여가시간, 숙달목적 지향성 학습습관, 학교 체육시간 관련 문항이 있었다. 본 연구는 국내 사회과학 패널 자료에 group lasso를 적용한 최초의 연구일 것으로 사료된다.

주제어: 한국아동청소년패널조사, 삶의 만족도, group lasso, 별점 회귀모형, 숙달목적 지향성 학습습관, 기계학습

* 한국교원대학교 교육학과 부교수, jeyoo@knue.ac.kr

** 한국교원대학교 교육학과 박사과정

I. 서 론

삶에 대한 만족, 행복은 인간이 일생을 살면서 추구하는 가장 큰 가치 중 하나이며, 생애 주기 중 어느 한 순간에 정지시키거나 유예시킬 수 없는 것이다(이상록, 김은경, 윤희선, 2015). 그러나 우리나라 어린이·청소년의 주관적 행복지수는 해당 지표가 있는 OECD에 가입된 23개국 중 최하위로 매우 낮은 수준이다(정은주, 2016). 특히 중학교에서 고등학교로 학교급이 높아지면서 삶에 대한 만족도가 낮아지는 추세가 보고된 바 있다(김소영, 윤기봉, 2016). 청소년기의 경험이 이후 전 생애에 걸쳐 지속적인 영향을 미칠 수 있다는 점을 고려할 때, 청소년의 삶의 만족도를 높이기 위한 범사회적인 관심과 노력이 필요하다. 청소년의 삶의 만족도 실태 조사 및 결정 요인을 탐색하고 청소년기 삶의 만족도가 이후 발달단계에 미치는 영향을 분석하는 연구가 활발하게 진행되고 있다(이상록, 김은경, 윤희선, 2015).

본 연구는 그러한 연구 중 한국아동·청소년패널조사(Korean Children and Youth Panel Study: 이하 KCYPS) 자료를 활용한 삶의 만족도(또는 행복감) 연구에 초점을 맞추어 분석하였고, 다음과 같은 특징을 파악할 수 있었다. 첫째, 선행연구는 기존 이론에 근거하여 KCYPS 자료의 일부 변수를 모형에 이용하였다. 둘째, 선행연구들은 비별점(non-penalized) 회귀분석 기법에 속하는 다중회귀모형, 분산분석, SEM(Structural Equation Model), HLM(Hierarchical Linear Model) 등을 주로 이용하여 삶의 만족도와 관련된 변인 간 관계를 모형화하였다. 그러나 비별점 회귀모형의 큰 단점 중 하나는 투입 변수가 많아질수록 회귀계수의 표준오차가 팽창하며 모형이 불안정해질 수 있다는 것이다. 회귀분석을 이용한 대다수의 선행연구들이 소수의 선택된 설명 변수만을 사용한 것은 이 때문이라고 할 수 있다. 셋째, 선행연구들은 모형 예측력(predictability)을 고려하지 않았다. 즉, 모형 구축에 쓰이지 않은 새로운 자료에 대하여 모형이 얼마나 잘 수행할지는 연구의 고려 대상이 아니었다.

한국청소년정책연구원의 KCYPS는 사백 개가 넘는 변수를 제공한다. 기존 연구에서 이 중 십여 개 정도의 변수만을 이용하는 것은, 방대한 KCYPS 자료를 충분히 활용하지 못하는 것이다. 비별점 회귀분석에서 수백 개가 되는 KCYPS 변수를 한 모형에 투입하는 것이 현실적으로 불가능하다. 반면, 최근 관심의 대상이 되고 있는 기계학습 기법을 이용한다면, 한 모형에 수백 개의 설명변수를 모두 투입하여 그 중 중요한 변

수를 선택할 수 있다. 그 결과 새롭게 발견된 변수가 있다면, 이는 기존 이론을 정비하고 보완하는 것에도 큰 도움이 될 것이다. 본 연구는 기계학습 기법 중 벌점 회귀 모형인 group lasso를 통하여 KCYPS 자료가 제공하는 수백 개의 변수들을 최대한 활용함으로써 중학생의 삶의 만족도에 영향을 미치는 변수를 탐색하고, 모형 예측력에 초점을 맞추어 모형을 구축하고자 하였다.

II. 선행연구 고찰

1. 개관

삶의 만족도는 ‘즐겁고, 희망적이며, 스트레스와 고민이 적은 상태로 스스로 느끼는 행복한 정서적 상태’라는 뜻의 ‘행복감’(한국청소년개발원, 2005)과 혼용되어 쓰인다. KCYPS에서는 삶의 만족도를 한국청소년개발원(2006)에서 연구된 세 개의 문항인 ‘나는 사는 게 즐겁다’, ‘나는 걱정거리가 별로 없다’, ‘나는 내 삶이 행복하다고 생각한다’의 세 개의 리커트 척도 문항을 이용하여 측정하고 있다.

RISS에서 “KCYPS”와 “삶의 만족도” 또는 “행복감”을 주제로 검색하였을 때 2016년 9월을 기준으로 열네 편의 학술논문을 찾을 수 있었다. 각 논문에서 이용된 연구방법을 중복하여 세었을 때, 가장 많이 사용된 연구 방법은 ANOVA를 포함한 일반 회귀분석으로, 총 아홉 편의 연구가 중다회귀분석, 단계적회귀분석, ANOVA 등을 이용하였다. SEM(Structural Equation Modeling), HLM(Hierarchical Linear Model)이 각각 세 편의 연구에서 이용되었으며, 군집분석을 이용한 논문이 두 편 있었다. LGM(Latent Growth Model), 의사결정나무(decision tree), 교차분석은 각각 한 편의 연구에서 사용되었다.

연구에 주로 쓰인 변수는 부모의 양육방식, 학교적응, 또래애착 등이었다. 부모의 양육방식은 청소년의 삶의 만족도와 관련하여 가장 많이 이용된 변수로 선행연구 열네 편 중 열두 편의 연구에서 사용되었다. 두 번째로 빈번하게 이용된 변수는 학교적응과 또래애착으로 각각 여덟 편의 연구에서 찾아볼 수 있었으며, 자아존중감은 일곱 편, 성별은 다섯 편, 자아탄력성, 전체성적 만족도, 지역사회관, 가구소득, 보호자에

대한 변수는 각각 네 편의 연구에서 발견할 수 있었다. 건강, 가족구성, 형제자매에 관한 변수는 각각 세 편의 연구에서, 정서문제와 학습습관은 두 편의 연구에서 사용되었다. 한 편의 연구에서만 사용된 변수들은 자아정체감, 진로정체감, 학업성취, 부모-친구 관계, 다문화 수용, 지역크기, 핸드폰 의존, 연간비행피행경험 유무, 고등학교 유형이었다. 선행연구에서의 이용 빈도대로 연구 결과를 분석하면 다음과 같다.

2. 선행연구 분석

선행연구에서 가장 많이 사용된 변수는 부모 양육방식으로 총 열두 편의 연구에서 이용되었다. KCYPs에서 조사하는 부모의 양육방식은 감독, 애정, 비밀관성, 과잉기대, 과잉간섭, 합리적 설명, 방임, 학대로 총 여덟 가지 하위 범주로 구성된다. 그러나 차수에 따라 조사된 하위 범주가 달라 분석에 사용된 양육방식의 종류가 연구마다 상이하였다. KCYPs에서 제공하는 양육방식에 관한 21개 문항을 모두 사용한 연구는 김소영, 윤기봉(2016), 김신경, 박민경, 안지선(2014), 박주희(2016), 최영걸, 유성모(2012)이다. 손보영, 김수정, 박지아, 김양희(2012)는 애정, 감독, 합리적 설명이라는 바람직한 양육방식을, 권세원, 이해현, 송인한(2012), 최창용(2015)은 애정적 양육방식에 대한 문항을, 박가화, 이지민(2014)과 조형정, 윤지영(2015), 천희영, 송영주, 이미란(2014)은 부정적 양육방식인 방임과 학대에 대한 문항을, 정득, 이종석(2015)은 방임, 그리고 정소희(2016)는 과잉간섭과 과잉기대에 대한 문항을 사용하였다. 이렇게 연구마다 사용된 양육방식의 하위 범주는 상이하였으나, 연구 결과는 청소년이 부모의 양육방식이 긍정적이라고 인식할수록 삶의 만족도가 높고, 부정적이라고 인식할수록 만족도가 낮은 것으로 나타나 대체로 일치된 모습을 보였다(권세원 외 2012; 김소영, 윤기봉, 2016; 김신경 외, 2014; 박가화, 이지민, 2014; 박주희, 2016; 손보영 외, 2012; 정소희, 2016; 조형정, 윤지영, 2015; 천희영, 송영주, 이미란, 2014; 최영걸, 유성모, 2014; 최창용, 2015). 양육방식이 삶의 만족도에 통계적으로 유의한 영향을 끼치지 않는다고 보고한 연구는 열두 편 중 정득, 이종석(2015)의 연구 한 편뿐이었다.

선행연구에서 두 번째로 빈번하게 사용된 변수 중 하나인 학교적응은 학습활동, 학교규칙, 교우관계, 교사관계 네 개의 하위 범주로 이루어져 있으며, 교사관계에 대한 문항만 사용한 권세원 외(2012)를 제외한 일곱 편의 연구가 네개의 하위 범주를 모두

사용하였다. 대부분의 연구에서 학교적응은 직접적으로 삶의 만족도에 영향을 미쳤고(천희영 외, 2014; 정득, 이종석, 2015), 동시에 매개변수로서 부모의 양육방식이 삶의 만족도에 미치는 영향을 조절하였다(손보영 외, 2012). 또는 부모의 양육방식이 학교 적응과 삶의 만족도 간의 관계를 매개하기도 하였는데(김신경 외, 2014), 박가화, 이지민(2014)의 연구에서는 학교적응이 양육방식과 삶의 만족도를 매개하지 않는 것으로 나타났다. 하위범주에 대하여 자세히 살펴보면 조형성, 윤지영(2015)에서는 교사관계가, 최영걸, 유성모(2014), 정득, 이종석(2015)에서는 교우관계가 삶의 만족도에 가장 큰 영향을 미쳤다.

또래애착은 의사소통, 신뢰, 소외의 세 범주로 구성되어 있으며, 신뢰 범주만 사용한 안지혜 외(2016)와 김소영, 윤기봉(2016)을 제외한 여섯 편의 연구에서 세 범주가 모두 사용되었다. 또래애착은 아토피 피부염 청소년의 삶의 만족도에 대하여 연구한 최창용(2015)을 제외한 모든 연구에서 삶의 만족도에 통계적으로 유의한 영향을 미치고 있음이 확인되었다(정득, 이종석, 2015; 정소희, 2016; 조형정, 윤지영, 2015; 천희영 외, 2014). 권세원 외(2012)는 또래애착이 삶의 만족도에 영향을 미치는 자아존중감과 자아탄력성에 직·간접적인 영향을 미치고 있음을 밝혔으며, 잠재성장모형을 사용한 김소영, 윤기봉(2016)은 초기에는 신뢰, 소외가 청소년의 삶의 만족도에 영향을 미치며, 이후에는 소외감의 영향력이 커져 소외감을 느낌수록 삶의 만족도가 낮아진다고 보고하였다. 안지혜 외(2016)와 조형정, 윤지영(2015)은 회귀분석을 통하여 삶의 만족도에 영향을 미치는 요인으로 또래애착의 의사소통, 신뢰를 제시하였으며, 안지혜 외(2016)는 의사결정나무를 이용하여 또래애착 중 신뢰가 삶의 만족도를 예측하는 변인이라고 결론지었다.

자아존중감은 일곱 편의 연구에서 사용되었는데 모든 연구에서 자아존중감이 삶의 만족도에 미치는 영향이 통계적으로 유의함이 확인되었다(권세원 외, 2012; 안지혜 외, 2016; 정득, 이종석, 2015; 조형정, 윤지영, 2015; 천희영 외, 2014, 최창용, 2015). 자아존중감에 영향을 미치는 요인은 양육방식, 학교적응의 또래 및 교사관계, 지역사회 관계 등이며(권세원 외, 2012), 자아존중감은 자아탄력성과 함께 삶의 만족도에 비교적 지속적인 영향력을 미치는 요인으로서 개인 특성의 중요성을 다시 한 번 강조하였다고 볼 수 있다(김소영, 윤기봉, 2016).

다섯 편의 연구에서 이용된 성별의 경우 그 결과가 연구에 따라 다소 달랐으나, 전

반적으로 삶의 만족도를 설명하는 변인에서 남녀 간 차이를 확인할 수 있었다. 최영걸, 유성모(2014)의 연구에서는 남학생보다는 여학생의 삶의 만족도가 통계적으로 유의하게 높았으나, 김소영, 윤기봉(2016)의 연구에서는 남학생의 삶의 만족도가 높았다. 천희영 외(2014)의 연구에서 전체성적 만족도는 여학생의 삶의 만족도에만 유의한 영향을 미쳤고, 박가화, 이지민(2014)의 연구에서도 양육방식이 학교적응과 삶의 만족도를 조절하는 경로가 여학생에서만 유의하였으며, 손보영 외(2012)의 연구에서는 남학생, 여학생 모두에게서 자아탄력성이 삶의 만족도에 유의하게 영향을 미쳤다.

자아탄력성, 전체성적 만족도, 지역사회관, 가구소득, 가족 관련 변수가 네 편의 연구에서 사용되었다. 자아탄력성은 자아존중감, 자아정체감, 진로정체감과 함께 자아인식의 하위 범주이며, KCYPS의 설문문항은 ‘나는 갑자기 놀라는 일은 당해도 금방 괜찮아지고 그것을 잘 이겨낸다.’, ‘나는 평소에 잘 해보지 않았던 새로운 일을 하는 것을 좋아한다.’, ‘나는 매우 에너지가 넘치는 사람이다.’ 등의 7개 문항으로 이루어져 있다. 자아탄력성을 사용한 모든 연구에서 자아탄력성은 청소년의 삶의 만족도에 긍정적인 영향을 미치고 있었다(권세원 외, 2012; 김소영, 윤기봉, 2016; 손보영 외, 2012; 최영걸, 유성모, 2014).

전체성적 만족도는 조형정, 윤지영(2015)와 천희영 외(2014)에서는 청소년의 삶의 만족도와 유의한 관계를 맺고 있는 것으로 나타났으나, 한부모 가정 청소년의 삶의 만족도를 연구한 박주희(2016)와 의사결정나무를 사용한 안지혜 외(2016)의 연구에서는 그렇지 않았다.

지역사회관은 지역사회인식과 공동체 의식으로 이루어져 있으며 조형정, 윤지영(2015)에서만 지역사회인식, 공동체 의식 문항이 모두 사용되었고, 나머지 세 편의 연구에서는 지역사회인식에 대한 문항만 사용되었다. 권세원 외(2012), 조형정, 윤지영(2015), 최영걸, 유성모(2014) 연구에서는 지역사회인식이 삶의 만족도에 통계적으로 유의한 영향을 미쳤으나, 안지혜 외(2016)의 의사결정나무 모형에서는 삶의 만족도를 예측하는 변인으로 선택되지 않았다.

가구소득은 만원, 백만원 단위로 조사된 연간 가구소득이나 욕구대비 소득비(income-to-needs ratio)에 따라 구분된 상, 중, 하위계층으로 사용되었으며, 네 편의 연구에서 모두 통계적으로 유의하지 않았다(박주희, 2016; 정소희, 2016; 조형정, 윤지영, 2015; 최창용, 2015). 선행연구에서 이용된 보호자와 관련된 변수는 부모의 학력 수준, 부모

의 건강이나 삶에 대한 만족도이다. 부모의 학력 수준과 건강 만족도는 네 편의 연구에서 모두 통계적으로 유의하지 않았는데, 부모의 삶의 만족도가 자녀의 삶의 만족도에 미치는 영향의 경우 최영걸, 유성모(2014)에서 통계적으로 유의하였다.

건강, 가족구성, 형제자매 관련 변수는 각각 세 편의 연구에서 이용되었다. 건강은 본인의 건강상태에 대한 주관적인 평가를 내리거나 중복 질병의 유무를 표시한 변수다. 중복 질환의 유무는 삶의 만족도와 관련이 없었으나(최창용, 2015), 건강에 대한 주관적 평가는 삶의 만족도와 통계적으로 유의한 관련이 있었다(조형정, 윤지영, 2015). 특히 스스로를 건강하지 못하다 여기는 청소년들의 삶의 만족도는 낮은 것으로 나타났다(권세원 외, 2012; 최창용, 2015). 세 편 이하의 연구에서 이용된 변수에 대하여 알아보면, 가족구성은 모자가정이 삶의 만족도에 부정적인 영향을 미치는 것으로 나타났다(정소희, 2016), 형제자매 유무는 모든 선행연구에서 삶의 만족도와 통계적으로 유의한 관계가 없는 것으로 나타났다.

정서문제와 학습습관은 각각 두 편의 연구에서 사용된 변수로, KCYPS에서 정서문제는 주의집중, 공격성, 신체증상, 사회적 위축, 우울의 네 개의 하위 범주로 구성되어 있다. 박주희(2016)는 우울, 공격성, 사회적 위축에 대한 스물 한 개 문항을, 안지혜 외(2016)는 우울, 주의집중에 대한 문항을 사용하였는데, 두 편의 연구가 공통적으로 삶의 만족도에 통계적으로 유의한 영향을 미치고 있다고 밝힌 요인은 우울이었다. 학습습관은 성취가치, 숙달목적 지향, 행동통제, 학업시간관리의 네 개의 하위 범주로 구성되어 있다. 이유리 외(2015)는 전반적인 학습습관이 우수할수록 삶의 만족도가 높다고 밝혔으며, 안지혜 외(2016)는 회귀분석을 통하여 학습습관 중 행동통제가 중학생의 삶의 만족도와 관련이 있는 변수임을 보여주었다.

진로정체감, 자아정체감, 부모-친구 관계, 다문화, 지역크기, 핸드폰의존, 비행, 고등학교유형, 학업성취는 열 네 편의 선행 연구 중 각각 한 편에서만 사용된 변수들이다. 안지혜 외(2016)에서 진로정체감, 연간비행피해경험 유무가, 조형정, 윤지영(2015)에서 지역크기, 핸드폰의존, 부모-친구관계, 다문화 수용 변수가, 박주희(2016)에서 고등학교 유형, 정소희(2016)에서 학업성취 변수가 사용되었다. 이 중 청소년의 삶의 만족도에 통계적으로 유의한 영향을 미치는 변수는 조형정, 윤지영(2015)의 지역크기와 정소희(2016)의 학업성취였다. 즉, 대도시 및 중소도시 대비 읍, 면 지역에 살수록, 학업성취가 높을수록 청소년의 삶의 만족도가 높았다.

III. 연구방법

1. 자료

1) 자료 정리

KCYPS는 다단계집락표집 방식으로 추출한 전국의 초1, 초4, 중1 학생들을 2010년부터 7년 동안 추적 조사하는 종단연구 자료이다. 조사 영역은 신체, 지적, 사회정서 등의 열다섯 개 소범주로 이루어진 개인발달 영역, 가정환경, 친구관계, 지역사회 환경 등 스물두 개 소범주로 이루어진 발달환경 영역, 그리고 배경변인으로 구성되어 있다. 본 연구에서 사용하는 자료는 6차년도 초4 패널로 2015년에 중학교 3학년이 된 학생 2,378명이 476개 문항에 대하여 응답한 결과이다.

본 연구의 데이터 클리닝 과정은 다음과 같다. 우선, 학부모 또는 학생이 미응답한 관측치와 분석에 불필요하다고 판단된 ID, 학교급, 가중치, 응답여부 등의 변수를 삭제하였다. 결측 비율이 높은 변수를 모두 모형에 포함시키는 경우 완전제거(listwise deletion) 후 유효한 사례 수가 전체 사례의 5% 미만이었다. 따라서 결측이 응답의 30% 이상인 학업중단사유, 비행 연간행동경험/피해 횟수, 방과후교육을 받은/받지 않은 이유, 사이버비행 연간 경험횟수, 성인용 매체 몰입도, 체험활동 연간 참여시간, 팬덤 활동 빈도 등의 변수가 삭제되었다. 마지막으로 완전제거 후 자료는 295개 문항에 대한 남학생 841명, 여학생 798명의 총 1,639명(69%)으로 정리되었다.

2) 반응변수

반응변수로 쓰인 삶의 만족도에 대한 문항은 '1. 매우 그렇다'부터 '4. 전혀 그렇지 않다'까지의 4단계로 응답하는 리커트 척도이다. 대부분의 선행연구가 삶의 만족도에 대한 세 문항을 역코딩하여 평균이나 총점을 구하였다(김소영, 윤기봉, 2016; 김신경 외, 2014; 박주희, 2016; 이유리 외, 2015; 정득, 이종석, 2015; 정소희; 조형정; 최영걸, 유성모, 2014; 최창용). 본 연구는 세 문항의 평균값을 구한 뒤 중앙값인 2보다 작거나 같은 경우를 '1'로, 2보다 큰 경우를 '0'으로 코딩하였다. 코딩 결과 삶의 만족도가

‘1’인 학생은 1,075명(65.6%), ‘0’인 학생은 564(34.4%)명으로 삶의 만족도가 양호한 학생이 그렇지 않은 학생보다 약 두 배 더 많았다.

3) 설명변수

이분형 문항이었으므로 ‘0’과 ‘1’로 더미코딩된 설명변수는 성별(GENDER), 형제자매 유무(FAM1D), 이성친구 유무(FRN3A), 연간 전학경험 유무(EDU1A), 보호자 방과후교육 유무(EDU5A02), 비행 연간행동경험 유무(DLQ1A01~14), 비행 연간피해경험 유무(DLQ2A01~07), 사이버비행경험 여부(MDA3A01~06), 체험활동 연간 참여경험 유무(ACT1A01~09), 동아리활동 연간참여 경험 유무(ACT2A01~02), 해외방문/교육 연간경험 유무(ACT4A,B), 선호연예인/운동선수 유무(ACY6A)였다.

이에 덧붙여 응답 결과 분포가 극단적이고 하위범주의 세분화가 불필요하다고 판단되어 ‘0’또는 ‘1’로 재코딩한 변수는 가족구성(FAM1A), 부모구성(FAM1B), 용돈 주는 사람(FAM5B)이다. 가족구성은 ‘부모(아버지와 어머니)+자녀’의 경우 ‘1’로 나머지 ‘한 부모+자녀’, ‘(한)조부모+자녀’ 등은 모두 ‘0’으로 코딩하였다. 부모구성은 ‘친아버지+친어머니’를 ‘1’로 그 외 ‘친아버지만’, ‘친어머니만’, ‘친아버지+새어머니’ 등은 모두 ‘0’으로 코딩하였다. 용돈 주는 사람은 ‘부모’를 ‘1’로, 나머지 ‘할아버지, 할머니’, ‘친척’, ‘형제자매’ 등은 모두 ‘0’으로 코딩하였다.

그 외 더미변수로 코딩된 범주형 변수는 모두 7개로, 자택지역:시도(ARA2A), 주택형태(HOUSE), 표준직업분류:부친(JOB4A), 종사상지위:부친(JOB5A), 장래희망교육수준:학생(FUR1A01), 장래희망교육수준:보호자(FURA02), 희망 체험활동1순위(ACT1D01)였다. 이러한 7개의 범주형 변수에 대하여 각각 16, 4, 9, 3, 5, 5, 8개의 더미변수가 생성되었다. 범주형 변수를 더미변수로 바꾼 후, 연속형 변수와 더미변수를 합한 총 설명변수 개수는 338개였다.

4) 평가 기준 및 소프트웨어

별점 회귀모형으로 청소년의 삶의 만족도를 예측하고 모형의 정확도를 확인하기 위하여 1,639명의 자료를 7:3의 비율로 나누어 각각을 훈련 자료(training data)와 시험 자료(test data)로 이용하였다. 이 때 원자료에서의 삶의 만족도 1과 0의 비율이 훈련

자료와 시험 자료에서도 유지되도록 하였다. 그 다음 훈련 자료로 모형을 구축하였는데, 10-fold CV(cross-validation; 교차타당화)를 이용하여 group lasso에서의 벌점모수인 λ 를 결정하였다(Hastie, Tibshirani, & Friedman, 2009; 유진은, 2016). 마지막으로 시험 자료로 예측 정확도(prediction accuracy), 특이도(specificity), 민감도(sensitivity)를 구하여 모형을 평가하였다(Hastie et al., 2009).

본 연구는 통계 계산과 그래픽을 위하여 개발된 무료 분석 프로그램인 R을 사용하여 삶의 만족도에 관련된 변수들의 영향력을 분석하였다. 결측치 처리 및 변수 재코딩, 훈련 및 시험 자료 분리에 사용된 패키지는 car, caTools이며, group lasso 벌점 회귀모형 분석에는 grpreg 패키지가 사용되었다.

2. Group Lasso

비벌점 회귀모형(non-penalized regression model)이 비편향(unbiased) 추정치 중 MSE(mean squared error)가 작은 추정치를 찾는 반면, 벌점 회귀모형(penalized regression model)에서의 추정치는 편향(biased) 추정치이기는 하나, 추정치의 분산을 줄임으로써 MSE 자체가 작아지도록 한다(유진은, 2016; Hastie et al., 2009). 다시 말해, 벌점 회귀모형의 목적은 편향을 감내하는 대신 분산을 줄임으로써 비벌점 회귀모형에서보다 MSE가 더 작은 모형을 찾아내는 데 있다.

벌점 회귀모형 중 Tibshirani(1996)가 제안한 축소추정 방법인 lasso(least absolute shrinkage and selection operator; LASSO)는 어떤 변수 계수를 0으로 만듦으로써 변수 선택을 가능하게 하는 방법이다. 따라서 lasso를 비롯한 벌점 회귀모형은 변수 수가 사례 수보다 더 많을 수 있는 빅데이터 시대에 특히 각광받고 있으며, 통계학, 컴퓨터공학 등의 분야에서 활발하게 연구되어 오고 있다. 국내에서도 통계학과 컴퓨터공학은 물론이고, 의학, 금융, 농학 등의 다양한 분야에서 적용되어 왔으며, 최근 KYPS 자료 분석에도 이용된 바 있다(유진은, 2016). 더 자세하게 설명하자면, lasso는 OLS(ordinary least squares)와 같은 손실함수(loss function) 식에 L_1 이라 불리는 회귀계수의 절댓값 합($\sum_j |\beta_j|$)이 벌점 함수(penalty function)로 추가되어, 기준을 충족하는 어떤 변수 계수를 0으로 축소추정함으로써 그 변수를 모형에서 제외시킨다(Hastie et al., 2009). 이 때 모형

에 남은, 계수가 0이 아닌 변수들이 모형을 설명하는 설명변수가 된다.

그러나 lasso는 주택형태, 직업 등과 같이 더미코딩이 필요한 범주형 변수의 경우 적절하지 않다. 예를 들어 KCYPS의 경우 주택형태를 단독주택, 아파트, 연립주택, 다세대주택, 상가주택, 보호 및 양육시설, 기타의 일곱 개 범주로 자료를 수집하였다. 이 경우 주택형태는 총 여섯 개의 더미변수로 코딩될 수 있으며, 이 여섯 개의 더미변수가 하나의 범주형 변수로부터 도출되었다는 변수 집단의 구조적 특성이 축소 추정 시 고려되어야 한다. 그런데 lasso의 경우 여섯 개의 더미변수가 하나의 범주형 변수에서 왔다는 사실을 고려하지 않는다는 문제가 있다. 이러한 문제를 해결하기 위하여 Yuan과 Lin(2006)이 제시한 방법이 바로 group lasso다. Lasso와 비교 시 group lasso는 벌점 부분이 다르다. Lasso에서는 L_1 벌점을 쓰는 반면, group lasso는 L_1 에 L_2 벌점인 회귀계수의 유클리드 노름(Euclidean norm; $\sum_j \beta_j^2$) 제곱 부분이 추가된다(Friedman, Hastie & Tibshirani, 2010). 즉, group lasso의 벌점은 lasso의 L_1 벌점과 능형회귀(ridge regression)의 L_2 벌점이 섞인 형태이므로, group lasso는 lasso보다 더 포괄적인 모형이라 할 수 있다.

본 연구에서는 group lasso 기법 중 Breheny(2015)의 ‘group exponential lasso (gel)’를 벌점모형으로 이용하였다. Group exponential lasso는 범주형 변수 집단 수준 및 더미변수 수준에서 중요한 변수를 선택할 수 있으며, 사례 수가 충분히 클 때 비편향 추정치에 가까운 계수를 추정할 수 있다는 장점이 있다(Breheny, 2015). Group lasso는 lasso에서와 마찬가지로 손실함수와 벌점함수의 합으로 구성된 함수 Q 를 최소화하는 회귀계수를 추정한다(식 (1)). 이 때 벌점함수 부분인 $P(\beta|\lambda)$ 가 벌점모수인 λ 를 통하여 축소추정을 담당한다. 식 (2)에서 group exponential lasso의 벌점함수를 구체화하였다. $f(\cdot)$ 가 0부터 무한대까지의 범위를 지나는 오목함수인 경우, 벌점함수는 J 개의 범주형 변수의 구조적 특징을 고려하는 해를 산출한다. Group exponential penalty에서 이용되는 ‘exponential penalty’ 식은 (3)과 같다. 즉, 벌점률(penalty rate)이 지수적으로 소멸하며, $f(\theta)$ 의 도함수는 0에 점근적으로 접근할 뿐, 0이 되지는 않는다는 것을 알 수 있다. 이 때 τ 는 지수적 소멸률(exponential decay rate)에 대한 조율 모수(tuning parameter)가 된다.

$$Q(\beta) = L(\beta | y, X) + P(\beta | \lambda) \quad (1)$$

$$P(\beta | \lambda) = \sum_{j=1}^J f(\|\beta_{j1}\|_1 | \lambda) \quad (2)$$

$$f(\theta | \lambda, \tau) = \frac{\lambda^2}{\tau} \left\{ 1 - \exp\left(-\frac{\tau\theta}{\lambda}\right) \right\} \quad (3)$$

IV. 결 과

1. Group LASSO 벌점 회귀모형 구축

훈련 자료를 이용하여 벌점모수 값에 따라 변수들의 회귀계수가 어떻게 변화하는지 그림 1과 같이 확인할 수 있었다. 그림 1에서 가로축은 벌점모수(λ)이고 세로축은 그 때의 회귀계수 값이다. 각각의 선들은 더미변수를 포함한 338개 설명변수를 뜻한다. 벌점모수 값이 0에 가까워질수록 회귀계수가 최소제곱법 결과에 가까워지고, 벌점모수 값이 증가함에 따라 회귀계수가 0으로 수렴되고 있음을 알 수 있다.

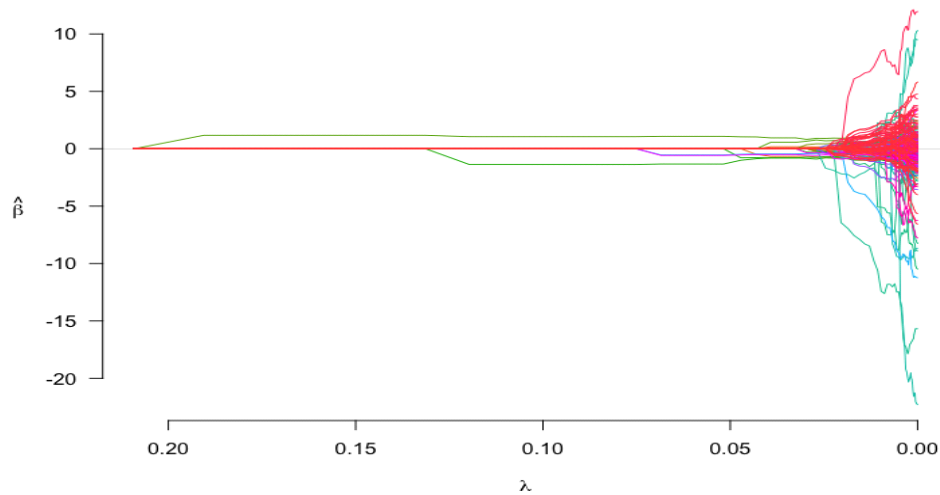


그림 1. 벌점모수와 회귀계수

최적의 벌점모수 값을 찾기 위하여 10-fold 교차타당화(cross-validation)를 실시한 결과는 그림 2와 같다. 가로축은 벌점모수에 자연로그를 취한 값이고, 세로축은 CV 오차(cross-validation error)이며, 수직으로 있는 점선은 CV 오차를 최소로 만들어주는 벌점모수 0.0357의 위치를 표시한 것이다. 이 값을 벌점모수 값으로 선택하여 벌점 회귀모형을 구축하였다. 이 벌점 회귀모형의 예측 정확도(prediction accuracy)는 0.7825였다. 즉, 새로운 관측치에 대하여 이 모형은 78.25%의 확률로 정확하게 예측하였다. 민감도(sensitivity)와 특이도(specificity)는 각각 0.6987과 0.8214였다. 교차타당화 시 이탈도(deviance)를 기준으로 하였다.

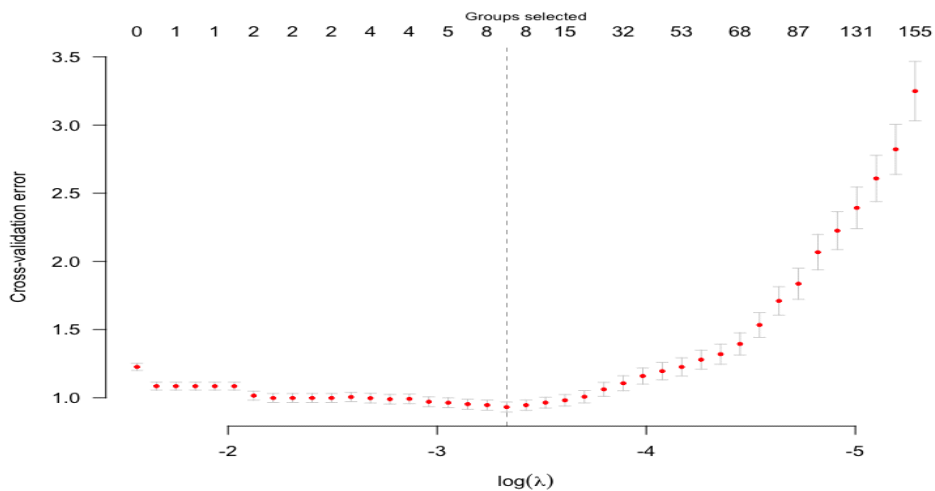


그림 2. CV 오차 기준 10-fold CV 결과

2. Group LASSO 벌점 회귀모형 분석 결과

벌점모수 값 0.0357을 이용하여 벌점 회귀모형을 구축하였을 때 회귀계수가 0이 아닌 설명변수는 모두 열다섯 개였다. 산출된 열다섯 개 설명변수의 변수명과 설명, 척도, 회귀계수는 [표 1]과 같다. 벌점 회귀모형에서는 정확한 편향을 측정하기 어렵기 때문에 회귀계수의 표준오차를 산출하지 않는 것이 일반적이다(유진은, 2016).

분석 결과, 중학교 3학년 학생의 삶의 만족도에 큰 영향을 미치고 있는 변수들은 자

아인식과 정서문제에 관련된 문항들이었다. 자아인식 중 자아존중감에 관련된 두 문항, ‘나는 나에게 만족한다’(PSY2A01)와 ‘나는 나에 대해 긍정적인 태도를 가지고 있다’(PSY2A10)에 긍정적으로 응답할수록 삶에 만족할 확률이 높았다. 정서문제에 관련된 문항에서는 우울에 대하여 묻는 ‘걱정이 많다’(PSY1E03), ‘불행하다고 생각하거나 슬퍼하고 우울해한다’(PSY1E02)와 신체증상 중 ‘자주 피곤해한다’(PSY1C06)가 선택되었다. 즉, 걱정이 많거나, 불행하다고 생각하거나 슬퍼하고 우울해하거나, 자주 피곤해할수록 삶에 불만족할 확률이 높았다. 자아존중감이 삶의 만족도에 긍정적인 영향을 미친다는 이러한 결과는 여러 선행연구들과 일치하는 결과이다(권세원 외 2012; 김소영, 윤기봉, 2016; 정득, 이종석, 2015; 조형정, 윤지영, 2015; 안지혜 외, 2016, 천희영 외, 2014; 최창용, 2015). 반면, 우울은 선행연구에서 적은 빈도로 사용된 변수임에도(박주희, 2016; 안주희 외, 2016), 본 연구에서는 두 번째로 큰 회귀계수를 보였다.

표 1

별점 회귀모형 분석 결과 회귀계수가 0이 아닌 15개 설명변수와 척도

순	변수명	변수 설명	척도	계수
1	PSY2A01	자아인식: 자아존중감-나는나에게 만족한다	Likert: 1(매우 그렇다) ~4(전혀 그렇지 않다)	-0.886
2	PSY1E03	정서문제: 우울- 걱정이 많다	Likert: 1(매우 그렇다) ~4(전혀 그렇지 않다)	0.821
3	PSY2A10	자아인식: 자아존중감-나는 나에 대해 긍정적인 태도를 가지고 있다	Likert: 1(매우 그렇다) ~4(전혀 그렇지 않다)	-0.781
4	INT1D	전체성적 만족도	Likert: 1(매우 만족한다) ~4(전혀 만족하지 않는다)	-0.471
5	FAM2F01	양육방식: 합리적설명-부모님(보호자)의 결정을 무조건 따르게 하기 보다는 왜 그래야 하는지 설명해주신다	Likert: 1(매우 그런 편이다) ~4(전혀 그렇지 않다)	-0.463
6	COM1A05	지역사회인식: 나는 우리 동네 사람들과 지내는 것이 좋다	Likert: 1(매우 그런 편이다) ~4(전혀 그렇지 않다)	-0.463
7	GENDER	성별	0(남자), 1(여자)	-0.409
8	PSY1C06	정서문제: 신체증상-자주 피곤해한다	Likert: 1(매우 그런 편이다) ~4(전혀 그렇지 않다)	0.336

순	변수명	변수 설명	척도	계수
9	PSY1E02	정서문제: 우울-불행하다고 생각하거나 슬퍼하고 우울해한다.	Likert: 1(매우 그런 편이다) ~4(전혀 그렇지 않다)	0.176
10	TIM8A	여가시간: tv/비디오 시청시간-등교일(시)	시간	0.140
11	INCOME1	가정 경제수준 평가	Likert:1(매우 잘 사는 편) ~7(매우 못사는 편)	-0.139
12	TIM5C	학업관련시간: 기타 공부시간-비등교일(시) (학교, 학원(과외) 숙제 이외의 공부시간)	시간	-0.105
13	TIM2A	학업관련시간: 학원/과외시간-등교일(시)	시간	-0.104
14	INT2B01	학습습관: 숙달목적 지향성-실수를 하더라도 무엇인가 배울 수 있는 어려운 내용을 좋아한다	Likert: 1(매우 그런 편이다) ~4(전혀 그렇지 않다)	-0.104
15	PHY3C	건강: 체육시간 중 운동시간	1(없다), 2(1시간)~5(4시간 이상)	0.065

* 변수는 회귀계수의 절댓값 크기 순으로 정렬하였다.

양육방식 중에서는 합리적 설명 관련 문항인 ‘부모님의 결정을 무조건 따르게 하지 않고 왜 그래야 하는지 설명해주신다’(FAM2F01)가 선택되었다. 즉, 양육방식이 합리적일수록 삶의 만족도가 높았다. 이는 김신경 외(2014), 최영걸, 유성모(2014)의 연구결과와 일치한다.

그 외 선행연구에서 다룬 변수 중 본 연구에서 선택된 변수들에는 전체성적만족도(INT1D), 지역사회인식(COM1A05), 성별(GENDER), 가정경제수준 평가(INCOME1) 등이 있었다. 전체성적 만족도는 조형정, 윤지영(2015) 및 천희영 외(2014)의 연구결과와 마찬가지로 자신의 전체 성적에 만족할수록 삶의 만족도가 높았다. 지역사회인식 변수에 대해서는 ‘나는 우리 동네 사람들과 지내는 것이 좋다’고 응답할수록 삶의 만족도가 높은 것으로 나타났다. 이는 권세원 외(2012), 조형정, 윤지영(2015), 최영걸, 유성모(2014)의 연구와 일치하는 결과이다. 또한 여학생보다는 남학생의 삶의 만족도가 높은 것으로 나와 김소영, 윤기봉(2016)의 연구결과와는 일치하였으나, 최영걸, 유성모(2014)의 연구결과와는 불일치하였다. 가구소득 변수를 사용한 네 편의 선행연구에서 가구소득은 유의한 변수가 아니었으나, 본 연구에서는 학생이 평가하는 가정 경제 수준 변수가 선택되었다. 즉, 학생이 본인 가정이 잘 사는 편이라고 인식할

수록 삶에 만족할 확률이 높았다.

기존 선행연구에서 다루어지지 않았으나 본 별점 회귀모형에서 선택된 변수에는 기타 학습시간, 여가시간, 학습습관-숙달목적 지향, 건강-체육시간 중 운동시간이 있었다. 등교일에 TV/비디오 시청(TIM8A)을 많이 할수록, 비등교일의 기타 공부시간(TIM5C)이 적을수록, 그리고 등교일의 학원/과외 시간(TIM2A)이 적을수록 삶의 만족도가 높았다. 또한, 학습습관 문항 중 숙달목적 지향성 관련 문항인 '실수를 하더라도 무엇인가 배울 수 있는 어려운 내용을 좋아한다'(INT2B01)에 그렇다고 답할수록 삶의 만족도가 높았다. 이유리 외(2015)에 사용된 학습습관은 네 가지 하위범주의 총점으로서 학습습관이 전체적으로 우수할수록 삶의 만족도가 양호하였다. 그러나 본 연구에서는 학습습관 중 특별히 숙달목적 지향성이 삶의 만족도와 통계적으로 유의한 것으로 나타난 것이 구별되는 점이다. 마지막으로 체육시간 중 운동시간(PHY3C)이 많을수록 삶에 만족할 확률이 높았다. KCYPS의 두 가지 건강 관련 문항 중 선택된 변수인 체육시간 중 운동시간은 열다섯 개 회귀계수 중 회귀계수가 가장 작았다.

V. 논 의

1. Group lasso의 변수 선택 결과

본 연구의 group lasso 결과 선택된 열다섯 개의 변수 중 자아인식, 정서문제, 양육방식, 전체성적 만족도, 지역사회인식, 성별, 가정경제수준 평가 관련 문항은 선행연구에서 이용된 바 있다. 본 연구 결과를 정리하자면, 학생들이 자신에게 만족하고 긍정적인 태도를 가질수록 삶에 만족할 확률이 높았다. 반면, 걱정이 많을수록, 불행하다고 생각하거나 슬퍼하고 우울해질수록, 그리고 자주 피곤해질수록 불만족 확률이 높았다. 또한, 전체성적 만족도가 높을수록, 부모님의 결정을 무조건 따르게 하기 보다는 왜 그래야 하는지 설명해줄수록, 동네 사람들과 지내는 것이 좋다고 답할수록, 남학생일수록, 그리고 학생 본인의 가정이 잘 산다고 인식할수록 삶의 만족도가 높았다. KCYPS 자료를 이용한 선행연구에서 성별의 경우 삶의 만족도에 미치는 영향이

일관적이지 못했고, 가정경제수준의 경우 통계적으로 유의하지 않았다. 그러나 그 외 변수들은 그 회귀계수가 통계적으로 유의하였으며, 선행연구에서의 회귀계수 방향 또한 일치하였다.

반면, 선행연구에서 빈번하게 이용된 변수 중 또래애착, 학교적응, 자아탄력성, 보호자, 가족구성, 형제자매 유무 변수는 group lasso 결과 선택되지 않았다. 열네 편의 선행 연구 중 한 두 편에서 사용된 연간비행피해경험 유무, 학업성취, 진로정체감, 부모-친구 관계, 다문화 수용, 지역크기, 핸드폰 의존 변수의 경우에도 마찬가지였다. 선행연구에서 다루지 않았으나 별점 회귀모형에서 선택된 변수로 등교일 학원/과외 시간 및 여가시간, 비등교일 공부시간, 숙달목적 지향성 학습습관, 학교 체육시간 문항 등이 있었다. 즉, 우리나라 중학생들은 공부시간이 적고, 실수하더라도 어려운 내용의 학습을 선호할수록, 학교 체육시간 중 운동시간이 길다고 인식할수록 삶의 만족도가 높았다.

등교일 학원/과외 시간 및 여가시간과 비등교일 공부시간 변수들은 그 자체로는 설명변수로서의 의의가 적을 수 있다. 학생들의 삶의 만족도를 높이기 위하여 비등교일 또는 등교일 공부시간을 줄이도록 권고하는 것이 현실적으로 가능한 것인지가 첫 번째 문제가 된다. 둘째, 공부시간을 줄이는 경우 전체성적 만족도가 떨어질 가능성도 배제할 수 없다. 표준화 후 전체성적 만족도가 공부시간 변수보다 회귀계수도 약 5배 더 크기 때문에, 전체성적 만족도를 높이는 것이 공부시간을 줄이는 것보다 삶의 만족도에 더 큰 영향을 끼친다고 할 수 있다. 그러나 공부시간 변수들은 무려 338개 변수 중 선택된 것으로, 이러한 변수들을 다른 중요한 연구변수와 함께 모형에 투입함으로써 모형의 설명력 및 예측력을 높일 필요가 있다. 즉, 공부시간을 비롯한 본 연구에서 새롭게 발굴한 변수들을 향후 삶의 만족도 연구모형에 투입하여 연구변수의 추정치 정교화에 이용하는 것이 좋다.

본 연구에서 새롭게 도출된 숙달목적 지향성 관련 학습습관 변수는 최근 강조되고 있는 과정중심평가와 맥을 같이 한다고 볼 수 있다. 2015 개정 교육과정에서 확정·고시한 과정중심평가에서는 학생의 학습과 평가를 수업과 연계하여 수업 중 이루어지는 각종 활동의 평가를 장려한다(교육부, 2015). 즉, 과정중심평가에서는 학습 결과만이 아닌, 학습 과정 그 자체가 목표가 된다. 학생들은 본인의 수준보다 어려운 과제에 도전하면서도 또래 학생이나 교사의 피드백을 통해 성공 경험을 쌓을 수 있는

것이다. 따라서 2016년 현재 17개 시도교육청에서 추진하고 있는 과정중심평가는 숙달목적 지향성 학습습관과 연결되어 학생들의 삶의 만족도에도 긍정적인 영향을 미칠 수 있을 것으로 사료된다. 학교뿐만 아니라 가정에서도 결과보다 과정을 중시하며, 자녀들이 실수하는 것을 두려워하지 않고 본인의 수준보다 좀 더 어려운 과제에 도전할 수 있도록 장려할 필요가 있다.

학교 체육시간 변수 또한 본 연구에서 처음으로 발굴한 변수로, 학교 체육시간 중 운동시간이 얼마나 되는지 학생들의 인식을 기반으로 측정된 것이다. 체육 시수는 교육과정에서 정해져 있으나, 일선 학교에서 여러 이유로 체육시간을 자유시간, 자습시간 등으로 대체하는 경우가 많다(조원표, 2016). KCYPS 자료는 아니지만, 2009년 청소년 건강행태 온라인조사 자료를 분석한 유진영, 김기만(2013)의 연구에서 신체활동이 수반된 체육시간이 길수록 스트레스 정도가 낮아지고 행복감이 높았다. 본 연구에서도 학생들이 인식하는 체육시간 중 운동시간이 길수록 삶의 만족도가 높았다. 따라서 학교 체육시간이 실질적으로 운영되고 있는지 점검하고, 최소한 교육과정에서 정해져 있는 체육 시수는 지키며, 특히 체육시간 중 실제 운동시간을 늘릴 수 있도록 교육청 및 각급 학교에 권고할 필요가 있다.

앞서 언급된 가정경제수준 관련 문항을 더 자세하게 살펴 보자면, KCYPS는 가구 연간 소득(INCOME) 변수를 1차부터 6차까지 보호자 응답으로 조사하며 5차부터 청소년 응답인 가정경제수준 평가(INCOME1) 문항을 추가하였다. 선행연구에서 이용되었던 가구 연간소득 변수는 부모(또는 보호자)가 응답하는 변수로 통계적으로 유의하지 않았으나, 본 연구의 가정경제수준 평가 변수는 청소년이 응답하는 변수로 group lasso 모형에서 선택되었다. 연구 방법이 다르므로 직접적인 비교는 어려우나, 가정경제수준 평가에 있어서는 부모의 응답보다 청소년의 직접 응답이 청소년의 삶의 만족도와 더 연관이 있는 것으로 보인다. 부모와 자녀의 가정경제수준에 대한 인식 차이가 청소년의 삶의 만족도에 미치는 영향이 향후 흥미로운 연구 주제가 될 수 있을 것이다.

2. 연구 의의 및 주의사항

청소년정책연구원의 KCYPS 자료는 삶의 만족도를 포함한 수백 개가 되는 변수들을 제공한다. KCYPS 자료를 이용하여 ‘삶의 만족도’에 어떤 변수들이 연관되어 있는

지를 알아보려는 선행 연구들은 그 중 십여 개의 변수 또는 요인을 선정하여 모형화해 왔다. 그러나 복잡다단한 인간을 대상으로 하는 사회과학 연구는 기존 이론에 근거한 몇 가지 변수 또는 요인만을 이용하는 데서 벗어나, 다른 새로운 변수들을 발굴할 필요가 있다. 특히 KCYPS가 제공하는 수백 개에 이르는 변수들은 그 이론적·실제적 타당성을 다수의 전문가들이 연구 설계에서부터 고려하여 수집된 변수들이다. 이러한 질 높은 변수들을 모두 한 모형에서 고려할 수 있다면, 선행연구에서 간과된 변수들을 찾아냄으로써 기존 이론을 정비할 수 있다. 또한, 본 연구는 KYPS 자료 분석에 lasso를 이용한 유진은(2016)의 연구에서 한 걸음 더 나아간 것이다. 즉, KCYPS에서 수집되는 범주형 변수의 구조적 특성을 견지하며 변수가 선택되도록 group lasso 기법을 적용하였다는 데 본 연구의 학술적 연구 의의가 있다.

새로운 방법론인 (group) lasso의 경우 기존 연구에서 이용되어 왔던 비벌점 회귀모형(non-penalized regression)과의 차이점을 주의해야 한다. 비벌점 회귀모형에서는 모형 선택 시 회귀계수의 통계적 유의성이 중요한 기준으로 이용되는 반면, 벌점 회귀모형에서는 회귀계수의 통계적 유의성이 아닌 모형 예측력(predictability)에 초점을 맞춘다. 즉, 모형 구축에 이용된 자료(훈련 자료)가 아닌 새로운 자료(시험 자료)에 대한 예측을 높이는 것이 벌점 회귀모형에서의 중요한 기준이 되는 것이다.

본 연구 결과, group lasso 모형에 포함된 15개 변수(문항)의 예측 정확도(prediction accuracy), 민감도(sensitivity), 특이도(specificity)는 각각 0.7825, 0.6987, 그리고 0.8214였다. 선행연구에서 자주 이용되었던 자이론중감(10 문항), 학교적응(20 문항), 지역사회(10 문항), 또래애착(9 문항), 긍정/부정 양육방식(29 문항), 성별(1 문항), 수입(1 문항), 성적만족도(1 문항), 부모변수(2 문항)를 필요한 경우 역코딩 후 평균을 구하여 모두 시험 자료에 투입한 결과, 예측 정확도, 민감도, 특이도는 각각 0.709, 0.603, 0.746였다. 즉, 기계학습 기법을 이용하여 KCYPS 자료의 15개 문항만으로 구축된 본 연구 모형의 예측 정확도, 민감도, 특이도가 기존 이론에 근거하여 83개 문항을 모두 이용하여 10개 변수로 구성된 모형의 예측 정확도, 민감도, 특이도보다 약 10%p 더 높았다.

다시 강조하자면, 벌점 회귀모형에서는 통계적으로 유의한 회귀계수를 선택하는 것이 아니라, 모형 예측력을 최대화하는 방향으로 모형 선택이 이루어진다(T. Hastie, personal communication, Jan. 28, 2017). 또한, 모형 예측에 기여하는 변수가 통계적

으로 유의하지 않을 수 있다(R. Tibshirani, personal communication, Jan. 28, 2017). 따라서 본 연구는 상관관계 또는 결정계수를 보고하지 않았으며, 이를 기존 연구 모형과 비교하지도 않았다. 대신, group lasso를 적용하여 얻은 15개 문항만으로 구성된 본 연구모형이, 기존 이론에 근거하여 83개 문항들을 이용한 모형보다 예측 정확도, 민감도, 타당도에 있어 약 10%p 가량 더 높았다는 사실을 통하여 본 연구 모형의 수월성을 입증하였다.

참고문헌

- 권세원, 이해현, 송인한 (2012). 청소년 행복감에 관한 연구: 청소년탄력성모델 (Adolescent Resilience Model)의 적용. **한국청소년연구**, 23(2), 39-72.
- 교육부 (2015). **초·중등학교 교육과정 총론**. 교육부 고시 제 2015-74호[별책 1].
- 김선아 (2012). 남녀중학생의 사회적 관계(부모, 또래, 교사관계)와 행복감의 관계. **청소년 시설환경**, 10(4), 15-26.
- 김소영, 윤기봉 (2016). 잠재성장모형을 이용한 청소년의 삶의 만족도 변화추정. **한국청소년연구**, 27(2), 271-296.
- 김신경, 박민경, 안지선 (2014). 학교생활적응과 부모의 양육행동이 청소년의 삶의 만족도에 미치는 영향: 부모 양육행동유형의 조절효과 검증. **청소년복지연구**, 16(2), 163-182.
- 김혜원, 홍미애 (2007). 우리나라 청소년들의 삶의 질 정도와 설명요인 분석. **청소년학 연구**, 14(2), 269-297.
- 박가화, 이지민 (2014). 부모의 양육행동이 남녀 청소년의 삶의 만족도에 미치는 영향: 자아정체감과 학교생활적응의 매개효과. **한국가족복지학**, 19(4), 997-1013.
- 박주희 (2016). 한부모 가정의 청소년이 지각한 부모양육태도 및 정서적 문제가 삶의 만족도에 미치는 영향. **한국가족자원경영학회지**, 20(1), 1-22.
- 손보영, 김수정, 박지아, 김양희 (2012). 바람직한 부모 양육태도가 초기청소년의 삶의 만족에 미치는 영향: 학교생활 적응 및 자아탄력성의 매개효과, 성차비교. **한국청소년연구**, 23(1), 149-173.
- 안지혜, 임희석, 윤유동 (2016). 한국아동·청소년패널조사 데이터를 이용한 중학생 삶의 만족도 분석. **디지털융복합연구**, 14(2), 197-208.
- 유진영, 김기만 (2014). 우리나라 중학교 청소년의 신체활동이 정신건강에 미치는 영향: 2013년 청소년건강행태온라인 조사를 중심으로. **디지털융복합연구**, 12(11), 395-405.
- 유진은 (2016). 데이터 마이닝 기법을 통한 교육 패널데이터 분석: 별점 회귀모형과 KYPs 자료. **아시아교육연구**, 17(3), 1-19.
- 이상록, 김은경, 윤희선 (2015). 청소년의 행복감이 청소년 발달에 미치는 영향. **사회과학연구**, 26(4), 413-438.
- 이유리, 박은정, 이성훈 (2015). 중학생의 학습습관 유형에 따른 학업성취와 삶의 만

- 족도의 차이. **학습자중심교과교육연구**, 15(11), 621-641.
- 정득, 이종석 (2015). 청소년 자아존중감과 환경특성이 삶의 만족도에 미치는 영향. **한국콘텐츠학회논문지**, 15(9), 273-284.
- 정소희 (2016). 부모의 과잉간섭과 과잉기대가 청소년의 정신건강에 미치는 상대적 영향: 우울, 공격성, 자존감, 삶의 만족도를 중심으로. **한국가족복지학**, 52, 313-342.
- 조원표 (2016. 10. 24.). '쉬는 시간' 전략한 체육수업. **한국교육신문**, p. 7.
- 조형정, 윤지영 (2015). 초등 고학년의 삶의 만족도 영향요인에 관한 연구. **순천향인문과학논총**, 34(1), 167-197.
- 천희영, 송영주, 이미란 (2014). 초등학교 5학년 아동의 삶의 만족도 관련 변인의 유형화 연구. **아동학회지**, 35(3), 71-92.
- 최영걸, 유성모 (2014). 초·중학생의 양육환경, 자아탄력성, 학교생활적응이 삶의 만족도에 미치는 영향. **뇌교육연구**, 14, 121-145.
- 최창용 (2015). 아토피 피부염 청소년의 삶의 만족도에 영향을 미치는 요인. **한국청소년연구**, 26(1), 111-144.
- 한국청소년개발원 (2005). **한국청소년개발지표 연구 II: 측정도구 개발을 중심으로** (연구보고05-R13). 서울: 한국청소년개발원.
- 한국청소년개발원 (2006). **한국청소년발달 지표조사 I: 결과부문 측정지표 검증** (연구보고06-R02). 서울: 한국청소년개발원.
- Brehereny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3), 731-740.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). *A note on the group lasso and a sparse group lasso*. Retrieved from <http://www.cs.odu.edu/~sji/classes/dm-2013s/papers/sparse-grlasso.pdf>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of statistical learning* (2nd ed.). New York: Springer.
- Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 68(1), 49-67.

ABSTRACT

Predictor exploration via group lasso: Focusing on middle school students' life satisfaction

Yoo, Jin Eun* · Rho, Minjeong**

This study explored predictors affecting KCYPS middle school students' life satisfaction levels, via penalized regression, specifically the group lasso. Among 338 KCYPS predictors after data cleaning had been conducted, a total of 15 were selected in the group lasso model. The prediction accuracy was determined to be 78.25%. Self-concept, depression, parental rearing styles, overall academic satisfaction, community recognition, gender, and home economic status were found to have an effect on students' life satisfaction, consistent with the findings of previous research. Study hours, mastery goal orientation, and students' perceived physical education hours were newly explored predictors, not yet investigated in previous research. As one of the first group lasso studies making use of social science panel data, this study successfully included categorical predictors as well as continuous ones in one group lasso model.

Key Words: KCYPS, life satisfaction, group lasso, penalized regression, mastery goal orientation, machine learning

투고일: 2016. 12. 5, 심사일: 2017. 1. 25, 심사완료일: 2017. 2. 14

* Associate Professor, Korea National University of Education, jeyoo@knue.ac.kr

** Doctoral graduate student, Korea National University of Education