아시아교육연구 17권 3호
Asian Journal of Education
2016, Vol. 17, No. 3, pp. 1-19.

데이터 마이닝 기법을 통한 교육 패널데이터 분석: 벌점회귀모형과 KYPS 자료

유진은(兪鎭銀)*

논문 요약

명확한 기존 이론이 없어도 축적된 데이터 분석을 통하여 결과를 도출할 수 있는 데이터 마이닝 기법이 빅데이터 시대에 각광을 받고 있다. 수렴 또는 과적합 등의 문제로 인해 소수의 변수만을 모형화해 온 기존 연구방법과 달리, 데이터 마이닝 기법으로는 수백 개의 변수를 한 모형에 투입할 수 있으며, 따라서 연구방법적 측면에서 여러 장점을 가진다. 국가기관에서 십수년 간 수집해 온 교육 패널자료는 양적 · 질적인 측면에서 데이터 마이닝 기법 적용에 적절하다. 본 연구는 빅데이터 분석에서 자주 이용되는 벌점회귀모형인 LASSO를 KYPS 5차 자료 분석에 이용함으로써 데이터 마이닝 기법의 교육 패널자료 적용 사례를 제시하였다. 수십 개의 변수만을 이용하였던 기존 연구와 달리, 본 연구는 총 315개의 설명변수를 한 모형에 투입하여 15개의 변수를 선택하였다. 기존 연구에서 모형화된 변수뿐만 아니라 새로운 변수를 발굴할 수 있었다. 본 연구의 함의 및 후속 연구 주제 또한 논의되었다.

주요어 : 교육 패널자료, 데이터 마이닝, 빅데이터, 벌점회귀모형, 한국청소년패널조사

Ⅰ. 서론

기존 HLM(Hierarchical Linear Modeling), SEM(Structural Equation Modeling), LGM(Latent Growth Modelling) 등과 같은 양적연구방법을 적용하는 연구에서는 이론이 중시된다. 즉, 이론

* 한국교원대학교 제1대학 교육학과 교수

의 토대 하에서 유의할 것으로 보이는 변수를 선택하고, 선택된 변수를 수집하고 통계모형에 투입한 후, 통계적으로 유의한 변수를 해석하는 것이 양적연구에서의 전통적인 절차다. 연구설계와 자료수집에 시간과 노력이 많이 들었던 과거에는 이러한 방식이 타당하였다. 그러나 소위 말하는 '빅데이터' 시대인 2016년 현재는 그에 맞는 데이터 마이닝 기법이 발전되어 오고 있다. 다시 말해, 빅데이터 시대에는 명확한 기존 이론에 의존하여 연구자가 직접 연구를 설계하여 자료를 수집하지 않고도 이미 축적된 자료를 데이터 마이닝 기법으로 분석하여 결과를 도출할 수 있다. 전통적 연구방법으로는 기존 이론을 벗어나기 쉽지 않기 때문에 새로운 변수 발굴이 제한적일 수밖에 없는 반면, 데이터 마이닝을 이용한다면 기존 이론에서 간과되어 왔던 새로운 변수를 찾아냄으로써 연구의 지평을 넓힐 수 있다.

새로운 변수 발굴 측면에서 볼 때, 국가기관에서 십수 년 간 수집해온 교육 패널자료가 데이터 마이닝 기법 적용 시 적절하다. 이를테면 한국청소년연구원의 한국청소년패널조사(KYPS), 한국직업능력개발원의 한국교육고용패널(KEEP) 등이 그러하다. 우선, 이러한 패널자료는 그 변수 수가 수백 개에 이르기 때문에 기존 연구에서 간과된 새로운 변수를 찾기 쉽다. 국가기관의 감독을 통해 자료의 질적인 측면이 담보되며, 표본 수 또한 수천 명에 이른다. 한국청소년연구원, 한국직업능력개발원, 한국교육개발원 등의 국가기관에서 매년 패널자료를 이용한 학술대회를 개최하고 다수의 연구자들이 호응해 오고 있다. 장시간 많은 비용과 노력을 들여 수집한 패널자료가 이렇게 소비되고 있다는 점은 매우 바람직하다. 그러나 대부분의 패널자료 연구가 HLM이나 SEM 등을 이용하고 있으므로 연구방법론 면에서 그다지 새롭지 않으며, 모수통계를 기반으로 하므로 연구에 제한점이 있는 것 또한 사실이다.

이러한 상황에서 새로운 방법론으로 부각되며 주목받고 있는 데이터 마이닝 기법을 교육 패널자료에 적용할 필요가 충분하다. 아직도 많은 학위논문이 타당화하기 힘든 준실험설계와 ANCOVA를 연구방법으로 이용하는 현실에서, 연구자가 패널자료에 데이터 마이닝 기법들을 이용하는 것만으로도 기존 연구와의 차별성을 담보할 수 있다. 또한, 데이터 마이닝 기법을 통한 연구 결과가 기존 방법보다 더 낫다면 이는 교육 관련 정책 수립 등에 이용되어 교육 관련 이해당사자인 학생, 학부모, 교직원, 정부 등에도 긍정적인 영향을 미칠 수 있을 것이다.

본 연구에서는 최근 각광받고 있는 벌점회귀모형 중 하나인 LASSO를 데이터 마이닝 기법으로 이용하였다. Ridge, LASSO, Elastic Net 등의 벌점회귀모형은 계수를 축소추정하는 특징으로 인하여 빅데이터 분석 시 주로 이용되는 기법이다. 연구자료로는 KYPS(Korea Youth Panel Study; 한국청소년패널조사)를 이용하였다. 최근 노동시장의 변화와 더불어 청년실업이 사회적 문제로 대두되며 청소년의 진로가 중요한 화두로 떠오르고 있는 점을 감안하여, 청소년의 진로선택과 관련 있는 설명변수를 탐색하고자 하였다. 정리하자면, 본 연구는 KYPS 5차 자료를 이용하여 데이터 마이닝 기법의 교육 패널자료 적용 사례를 제시하였다. KYPS 5차 자료를 이용한

이유는, 5차가 KYPS의 마지막 자료 수집 해로서 다수의 KYPS 연구들이 5차 자료를 이용하였으며, 2016년 기준 시 자료가 공개된 지 7년 이상 경과하였으므로 그동안 관련 연구가 충분히 출판되었을 것으로 사료했기 때문이다. 연구의 이론적 배경보다는 통계적 자료 분석을 통한 변수 추출이 주가 되는 데이터 마이닝 기법의 특성을 견지하며 분석을 먼저 수행하였고, 그 결과를 진로선택 관련 선행연구 결과와 비교·분석하였다.

Ⅱ. KYPS 선행연구

RISS와 Google Scholar에서 “KYPS 5차”를 주제어로 검색한 후, 마지막 해인 2008년 자료를 이용한 논문을 고르면, 2016년 5월을 기준으로 총 87 건의 학술논문을 찾을 수 있다. 총 87건의 논문 중 KYPS 5차 자료만을 이용한 논문은 14 편이었고, 나머지 73 편의 논문은 2년 이상의 KYPS 자료를 이용하여 분석한 논문이었다. 연구에 주로 쓰인 변수는 학교 적응, 스트레스, 진로결정 또는 진로성숙, 사교육, 부모애착, 교사애착, 또래애착, 자아존중감, 자기통제력, 가족구성형태, 가구소득, 학업성적, 부모 폭력 등으로 다양하였다. 연구 방법의 경우 다중회귀분석, SEM(Structural Equation Modeling), HLM(Hierarchical Linear Model), LGM(Latent Growth Model), ARCL(AutoRegressive Cross-Lagged model), 로지스틱 회귀모형 등의 회귀모형이 주를 이루었다.

KYPS 자료를 분석한 87 건의 학술지 논문 중 9 건이 진로성숙 또는 진로미결정을 반응변수로 하며 청소년의 진로에 초점을 맞추는 연구였다. 연구방법론으로 분류하자면 9 건 중 HLM 논문 두 편, SEM 논문 세 편, 그리고 ARCL과 LGM 논문이 각각 두 편이었다. 또 다른 특징으로, 9건의 연구에 이용된 여러 변수들이 문항묶음(item parceling)으로 구성된 점이 있다. KYPS의 많은 문항들이 1부터 5까지의 리커트(Likert) 척도로 측정되었는데, 9건 모두 몇몇 문항을 묶은 후 그 평균 또는 합으로 새로운 변수를 구성하여 모형에 투입하였다.

어윤경(2008)은 KYPS 중2 패널 1차~5차 자료를 이용하여 진로만족도와 진로미성숙의 관계를 선형 HLM과 이차함수 HLM으로 분석하였다. 진로관련 강연이나 수업, 소집단 활동, 적성검사, 상담, 직접체험 프로그램, 직업훈련, 책/잡지 열독이 어느 정도 도움이 되었다고 생각하는지를 알아보는 진로교육 경험 7문항의 총점을 진로만족도로 구성하였다. 종속변수인 진로미성숙 또한 7문항에 대한 총점으로, 향후 진로 관련하여 적성 및 소질을 잘 알지 못한다, 정보가 부족하다, 선택하기 힘들다, 진로가 자주 바뀐다, 부모님과 의견차이가 크다, 미리 설정해 봐야 아무런 소용이 없다, 나 자신보다 부모님의 의견을 따르는 편이다 등이 그 문항들이었다. 선형 HLM 결과, 진로미성숙은 학년이 올라갈수록 증가하다가 일정 시점에서 감소하기 시작하였다. 이차함

수 HLM 결과, 진로교육 만족도가 높을수록 초기치 진로성숙이 높고 진로성숙 하락폭이 작으며 이후 급격하게 진로가 성숙되는 것을 알 수 있었다.

김민선과 서영석(2010)은 KYPS 중학교 3학년 시기부터 고등학교 3학년까지의 2차~5차년도 자료를 이용하여 부모애착, 또래애착, 교사애착, 자기효능감, 부모갈등, 또래갈등, 학교장벽 등의 변수가 진로미결정에 미치는 영향을 HLM으로 분석하였다. 부모애착, 또래애착, 교사애착 변수를 위하여 각각 6, 5, 3개의 문항을, 자기효능감을 위하여 3개 문항을, 부모갈등, 또래갈등, 학교장벽 변수를 위하여 각각 4, 4, 2개 문항을 이용하였다. 이에 덧붙여 학생 성별, 부학력, 모학력을 배경변인으로 이용하였다. 반응변수인 진로미결정의 경우 김민선과 서영석(2010)은 앞선 어윤경(2008)의 7문항 중 '진로를 미리 설정해 봐야 아무런 소용이 없다'를 제외한 6문항의 평균을 산출하였다. 즉, 10가지의 설명변수를 위하여 총 30개 문항을 이용했으며, 하나의 반응변수를 위해 6문항을 이용하였다. 분석 결과, 남학생의 진로미결정이 높은 편이었고, 중3학생의 경우 높은 자기효능감이 낮은 진로미결정과 연관되어 있었다. 또한 부모갈등, 또래갈등, 학교장벽이 높을수록 진로미결정이 높았으며 그 중 또래갈등이 가장 영향이 컸다. 반면, 여학생이거나, 부모와의 갈등이 많거나, 학교장벽을 높게 지각할수록 진로미결정 감소율이 낮았다.

허성호와 정태연(2010)은 2003년부터 2008년의 KYPS 자료를 이용하여 중학생, 고등학생, 대학생의 진로의식 및 삶 만족에 관계성과 자아관이 얼마나 영향을 주는지 SEM을 통하여 분석하였다. 먼저 직업성숙도와 진로성숙도는 각각 관련된 7문항을 이용하여 외생 잠재변인인 진로의식을 측정하였다. 내생 잠재변인인 자아관을 측정하기 위하여 자아존중감, 자기신뢰, 자아낙인을 각각 6, 3, 2문항의 문항묶음으로 구성하였고, 역시 내생 잠재변인인 관계성을 측정하기 위하여 부모애착, 친구애착, 정서조절을 6, 3, 3 문항으로부터 문항묶음을 만들었다. 내생 잠재변인이자 외생 잠재변인인 일상생활은 일탈과 삶만족으로 구성되는데, 일탈의 경우 총 14가지(대학생은 12가지) 문항으로, 삶만족은 단문항으로 측정되었다. 허성호와 정태연(2010)은 다집단(multi-group) SEM을 이용하는 대신 각 학교급에 대해 각각 다른 SEM 모형을 만들었다. 중학생의 경우 외생 잠재변인인 관계성과 자아관이 일상생활에 영향을 미쳤으나 진로의식에는 영향이 없었다. 고등학생의 경우 관계성이 일상생활과 진로의식에 모두 영향을 주며 자아관은 일상생활을 매개변인으로 하여 진로의식에 영향을 준 반면, 대학생 모형에서 통계적으로 유의한 경로는 찾아볼 수 없었다.

김재철, 황매향과 김아영(2011)은 체험활동, 긍정적 자아관, 내적직업가치관이 진로성숙에 미치는 영향을 알아보기 위하여 다집단 SEM으로 2007년 고등학교 3학년 학생 KYPS 자료를 분석하였다. 외생 잠재변수인 체험활동은 진로활동, 수련활동, 자원봉사활동, 동아리활동으로 구성되며 각각 7, 6, 6, 8개 문항을 활용하였다. 매개변수로 이용된 긍정적자아관은 자아존중감, 자기신뢰감, 정서조절감의 문항묶음이었는데, 각각 6, 3, 3개 문항을 이용하였다. 역시 매개변수로 이

용된 내적직업가치관은 자율성, 헌신성으로 구성하였으며, 각각 4개의 문항을 이용하였다. 내생 잠재변수로 이용된 진로성숙은 진로결정성, 진로독립성으로 구성되며, 각각 4개와 3개의 문항을 이용하여 측정되었다. 성별을 집단으로 하는 다집단 SEM 결과, 형태동일성과 측정동일성이 충족되었다. 내적직업가치관과 긍정적자아관이 진로성숙에 직접영향을 미치고, 긍정적자아관은 내적직업가치관을 매개변인으로 하여 진로성숙에 간접영향 또한 끼치는 것을 알 수 있었다. 외생 잠재변수인 체험활동은 매개변수인 긍정적자아관과 내적직업가치관에 직접영향을 주었다. 후속 단계로 실시된 카이제곱 검정 결과, 여러 경로에서 통계적으로 유의한 남녀 차이를 찾아낼 수 있었다. 이를테면 체험활동이 진로성숙에 직접효과는 주지 못하였으나, 내적직업가치관을 매개로 진로성숙에 미치는 간접효과는 여학생에게 더 컸다. 반면, 체험활동이 긍정적자아관과 내적직업가치관을 차례로 매개하여 진로성숙에 미치는 영향은 남학생에게 더 컸다.

신선아와 전종설(2015)은 KYPS 중2 패널 4차 자료인 고등학교 2학년생에 대하여 친구애착, 부모애착, 교사애착, 그리고 자기효능감이 진로성숙에 어떤 영향을 보이는지 다집단 SEM을 이용하여 분석하였다. 내생 잠재변수인 진로성숙의 경우 7개의 문항을 이용하되, 문항묶음을 통해 세 개의 측정변수로 재구성하였다. 외생 잠재변수인 부모애착의 경우에도 마찬가지로 6개의 부모애착 관련 문항을 3개의 측정변수로 재구성하였다. 친구애착, 교사애착, 자기효능감의 경우 각각 3문항을 이용하였다고 했는데, 각 문항을 측정변수로 이용하였는지 아니면 문항들의 합이나 평균을 하나의 측정변수로 이용하여 분석하였는지는 불확실하다. 부가적으로 월평균 가구소득과 성적을 통제변수로 분석에서 이용하였다. 구조모형 분석 결과, 부모애착, 친구애착, 교사애착은 자기효능감을 매개변수로 하여 진로성숙에 정적인 영향을 미쳤으며, 친구애착, 교사애착, 자기효능감은 진로성숙에 직접효과를 보였다. 성별을 집단으로 하는 다집단 SEM 결과 부분측정동일성이 충족되었으나, 구조동일성 모형의 경우 성별 차가 통계적으로 유의하지 않았다. 구조모형에서 교사애착과 진로성숙이 부적인 관계를 보였는데, 여학생의 경우에 교사애착과 진로성숙의 부적인 관계가 통계적으로 유의했고 남학생의 경우에는 통계적으로 유의하지 않았다.

허균은 4편의 논문에서 진로성숙 또는 진로장벽을 반응변수로 하여 부모애착, 자아존중감, 진로경험활동 등의 설명변수의 영향을 ARCL(AutoRegressive Cross-Lagged model) 또는 LGM(Latent Growth Model)으로 KYPS 자료를 분석하였다. 2010년 논문에서는 중2 패널의 1, 3, 5차 KYPS 자료를 이용하여 진로경험활동과 진로성숙의 관계를 ARCL로 분석하였다. 진로성숙의 경우 KYPS에서 제공하는 7개 문항 중 부모 관련 문항 2개를 제외한 5개 문항을 세 개의 측정변수로 재구성하여 이용하였다. 진로경험활동은 진로관련 강연이나 수업, 소집단 활동, 적성검사, 상담, 직업훈련, 책/잡지 열독의 6개 활동의 여부를 합산하여 이용하였다. 3시점에서의 진로경험활동과 진로성숙으로 가능한 8개 계수는 모두 통계적으로 유의하였다. 다시 말해, 이전의 진로경험활동은 이후의 진로경험활동에 영향을 주며, 이전의 진로성숙 또한 이후의 진로성숙

에 영향을 주었다. 또한 진로경험활동과 진로성숙은 각각 교차지연 효과를 보여주었다.

허균(2012a)은 같은 KYPS 자료로 자아존중감과 진로장벽 간 관계를 같은 ARCL로 분석하였다. 자아존중감은 관련 6개 문항을 3개씩 묶어 긍정적 자아개념과 부정적 자아개념의 두 가지 측정변수로 재구성하였다. 허균(2012a)은 진로장벽을 5가지 하위개념으로 보고 각 하위개념에 해당하는 문항들을 선정하고 그 합 또는 평균을 측정변수로 이용하였다. 자기이해부족 및 진로직업 정보부족의 경우 2개 문항을, 성역할 고정관념의 경우 여성과 남성 문항 각각 3문항씩을 이용하였다. 중요한 타인과의 갈등, 미래에 대한 불확실성, 경제적 어려움은 각각 2, 1, 1개 문항으로 구성하였다. 연구 결과, 자아존중감과 진로장벽은 각각 통계적으로 유의한 순차효과를 보였고, 자아존중감의 진로장벽으로의 교차지연 효과는 모두 유의하였으나, 반대 방향의 교차지연 효과는 유의하지 않았다.

허균(2012b)은 중2 패널의 1차~5차 자료를 모두 이용하여 부모애착, 자아존중감, 성별이 진로성숙도에 미치는 영향을 연구하였다. 부모애착, 자아존중감, 진로성숙도 모두 KYPS가 제공하는 문항묶음인 6, 6, 7개 문항의 평균으로 측정하여 LGM(Latent Growth Model)으로 분석하였다. 연구 결과, 학년이 올라갈수록 자아존중감이 올라갔고, 초기시점에서 여학생의 진로성숙도가 높았으나 변화율에 있어 성별 차이는 없었다. 부모와의 애착이 클수록 진로성숙도가 높은 편이었으며, 부모애착과 진로성숙은 동시효과 및 지연효과가 모두 정적인 관계를 가지며 통계적으로 유의하였다. 허균(2013)은 패널을 바꿔 초4 패널의 1차~5차 자료를 이용하여 역시 LGM으로 분석하였다. 허균(2013)에서는 부모애착과 진로성숙만으로 모형을 만들었는데, 이 변수들의 측정방식은 허균(2012b)과 동일하였다. 허균(2012b)에서와 마찬가지로 부모애착과 진로성숙은 동시효과 및 지연효과가 모두 통계적으로 유의하였으며, 그 관계는 모두 정적이었다.

III. 연구 방법

1. 자료

사회과학 연구에서 널리 이용되고 있는 종단연구 자료 중 하나인 KYPS는 표집된 학생들을 매년 추적조사 하였다. 2003년 중학교 2학년생을 2008년 대학교 1학년이 될 때까지 6차에 걸쳐 조사한 자료와 2004년 초등학교 4학년을 2008년 중학교 2학년이 될 때까지 5차에 걸쳐 조사한 자료가 있다. KYPS는 층화다단계집락표집을 이용하여 먼저 15개의 지역별로 층화 후 집락표집으로 학교를 추출하고, 마지막으로 무선표집으로 학급을 추출하여 2003년 3449명의 중학교 2학년생과 2004년 2844명의 초등학교 4학년생을 표집하였다(Korea Youth Panel Survey, n. d.).

본 연구는 KYPS의 초등 5차 자료를 분석하였다. 이는 2008년에 중학교 2학년이 된 총 2844명의 학생들을 대상으로 683개 변수를 수집한 것이다. KYPS 자료의 단위 무응답(unit nonresponse)을 제거하기 위하여 설문지 응답여부가 '응답 안함'인 관측치들을 모두 삭제하였고, 항목 무응답(item nonresponse)을 줄이기 위하여 683개의 변수 중 30% 이상의 학생이 무응답한 변수들을 삭제하였다. 학생 ID, 설문지 응답여부 등과 같은 분석에 부적절한 변수 또한 제외하였다. 마지막으로 항목 무응답인 학생들을 삭제한 결과, 2055명(72%)에 대한 315개 변수로 자료가 정리되었다.

본 연구에서는 학생의 진로선택 여부를 반응변수로 하고, 나머지 314개 변수를 설명변수로 하였다. 즉, 314개 변수를 모두 이용하여 중학교 2학년생의 진로선택 여부를 분류하는 것이 연구 목적이었다. 진로선택 여부는 '귀하는 자신이 어른이 되어서 하고 싶은 직업을 정해 놓은 상태인가요?'를 물어보는 문항 q2w5에 대하여 생각해 놓은 직업이 있다고 답한 경우 '1', 아직 정해 놓은 장래의 직업이 없다고 답한 경우 '0'으로 코딩하였다. 자료 정리 후 문항 q2w5가 '0'인 학생은 436명(21%), '1'인 학생은 1619명(79%)이었다. 대부분의 KYPS 문항들이 리커트 식(Likert-like) 척도를 주로 이용했는데 분석의 편의 상 연속변수로 취급하였다. 문항묶음(item parcel)을 이용하지 않았으므로 모든 문항에 대하여 역코딩은 하지 않았다.

범주형 자료를 다루는 명명척도 변수의 경우 더미코딩으로 변환하여 이용하였다. 각 범주별로 빈도가 크게 차이 났기 때문에 가장 반응 빈도가 높은 한 범주 대 나머지 범주들의 합으로 문항을 재코딩하여 분석하였다. 예를 들어 중학교 졸업 후 향후 진로를 물어보는 문항(q2k1w5)은 무려 27개의 범주로 매우 자세하게 자료를 수집하려고 하였다. 그러나 27개 범주 중 7개의 범주에 답한 학생은 한 명도 없었으며, 나머지 11개 범주에도 50명도 채 안 되는 학생이 답하는 등 범주 간 빈도 불균형이 상당히 심각하였다. 따라서 '아직 구체적으로 정하지 못함'의 범주인 3, 13, 21, 24, 27을 하나로 묶어 0으로, 나머지를 1로 코딩하였다. 그 외 횟수나 시간 등을 측정하는 비율척도의 경우 변수 그대로 분석에 이용하였다.

2. 벌점회귀모형

벌점회귀모형은 계수에 벌점(penalization)을 부과함으로써 과적합(overfitting) 문제를 해결하고자 하는 축소추정법(shrinkage estimation methods)의 일종이다. 일반선형회귀(general linear regression)에서 이용되는 OLS(Ordinary Least Squares; 최소제곱법)는 불편 추정치(unbiased estimator) 중 분산(variance)이 최소인 추정치를 찾아낸다. 반면, 벌점회귀모형은 불편 추정치는 아니더라도 분산이 작은 추정치를 찾아냄으로써 전체 MSE(mean squared error; 평균제곱오차)를 낮출 수 있다. 즉, 벌점회귀모형을 통하여 OLS와 비교 시 MSE가 더 작은 추정치를 찾을 수

있다는 강점이 있다.

능형회귀(ridge regression), Elastic Net과 더불어 LASSO(Least Absolute Shrinkage and Selection Operator)가 벌점회귀모형의 대표적인 방법으로 꼽히며, 빅데이터 분석에서 즐겨 이용된다. 이 중 1996년 Tibshirani가 발표한 LASSO는 계수 추정 및 변수 선택을 동시에 수행한다는 장점으로 인하여 근래 기계학습(machine learning), 통계학습(statistical learning), 그리고 예측분석(predictive analytics) 등에서 각광받고 있다(Fan et al., 2014; Lang et al., 2014; Varian, 2014).

LASSO 추정식은 수식 (1)과 같다. N개의 표본에 대하여 P개의 설명변수 x로 반응변수 y를 추정할 때, λ가 벌점모수(penalty parameter)가 되며, λ값에 따라 축소추정을 어느 정도로 할 것인지 결정된다.

$$(1)\ \hat{\beta} = \arg min\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P}|\beta_j|\right\}.$$

식 (1)에서 λ가 클수록 계수가 0에 수렴을 하고, λ가 작아질수록 최소제곱법의 추정치와 가까워진다. 벌점모수, 즉 λ를 크게 하여 어떤 계수 추정값을 0으로 만드는 경우 자동적으로 변수선택이 이루어진다. 따라서 λ의 크기를 결정하는 것이 LASSO를 통한 계수 추정에 중요한 부분이 되며, 이 때 일반적으로 CV(Cross-Validation; 교차타당화)가 이용된다(Hastie, Tibshirani, & Friedman, 2009). 10-fold CV의 경우, 자료를 10 부분으로 나눈 후 돌아가며 자료의 $\frac{9}{10}$과 $\frac{1}{10}$을 각각 훈련자료와 시험자료로 이용하여 예측오차 평균을 구한다. 이를 모든 λ 값에 대해 반복하여 예측오차를 가장 작게 하는 λ 값을 찾게 된다.

LASSO와 같은 벌점회귀모형은 의학(김보현 외, 2015; 박철용, 계묘진; 2013), 금융(송상윤, 2015; 진슬기, 김광래, 박창이, 2011), 농업(박민수, 김태헌, 조은석, 김희발, 오희석, 2014) 등의 다양한 분야에서 근래 활발하게 적용되고 있다. 반면, 교육학을 포함한 사회과학 분야에서는 벌점회귀모형을 적용한 연구를 찾기 어려웠다.

3. 통계 프로그램 및 연구모형

SPSS version으로 입력된 데이터를 csv 파일로 변환하여 R로 불러들인 후, LASSO 추정을 위하여 R version 3.1.2의 'glmnet' 패키지를 이용하였다. 벌점회귀모형에서 벌점모수를 찾기 위한 교차타당화(CV; Cross-Validation)가 필수적이다. 본 연구에서는 전체 자료를 10 부분으로 나누

어 시험자료와 훈련자료로 이용하는 10-fold CV를 이용하여 벌점 모수를 선택하였으며, 이 때 이탈도(deviance)를 기준으로 하였다.

본 연구는 로지스틱 회귀모형이 연구모형이었다. 중학교 2학년생이 진로를 선택한 경우 집단 1, 진로를 선택하지 못한 경우 집단 0으로 코딩하여 로지스틱 회귀모형의 반응변수로 이용하였다. 연구모형은 식 (2)와 같다. 이 때 β^T는 314개 설명변수에 대한 회귀계수의 벡터다.

$$(2)\ \log\frac{P(G=1|X=x)}{P(G=0|X=x)}=\beta_0+\beta^T X$$

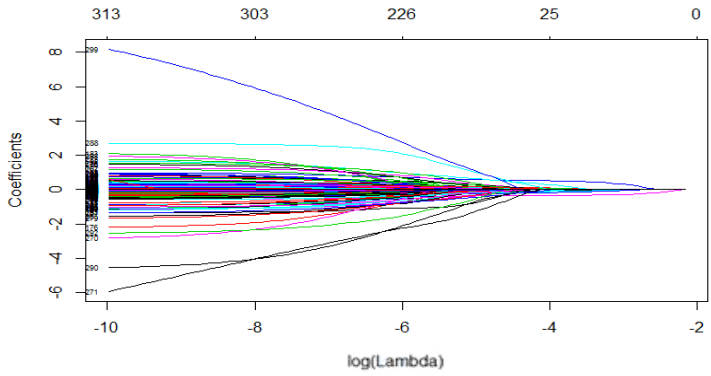
따라서 LASSO를 이용한 로지스틱 회귀모형은 식 (3)을 이용하여 추정된다(Hastie et al., 2009). LASSO는 변수의 척도에 민감하게 반응하므로 코딩이 끝난 모든 변수를 평균 0, 표준편차 1로 표준화한 후 분석에 이용하였다.

$$(3)\ \max_\beta\left\{\sum_{i=1}^{N}[(y_i(\beta_0+\beta^T X_i)-\log(1+e^{\beta_0+\beta^T X_i})]-\lambda\sum_{j=1}^{P}|\beta_j|\right\}$$

Ⅳ. 결과

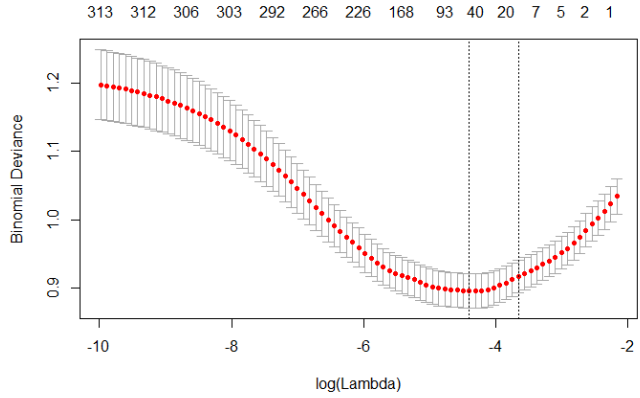
1. LASSO 벌점회귀모형 구축

[그림 1]에서 가로축은 벌점모수에 자연로그를 취한 값이고 세로축은 그 때의 회귀계수 값이다. 각각의 선들은 314개 설명변수의 회귀계수가 벌점모수(λ) 값이 증가함에 따라 0으로 축소되는 것을 보여준다. [그림] 1 상자의 맨 윗부분 숫자는 각 벌점모수 값에 대한 설명변수의 개수를 나타낸다.



[그림 1] 벌점모수와 회귀계수

다음 단계에서 벌점모수 값을 찾기 위하여 CV를 이용하였다. [그림 2]는 이탈도 기준 10-fold CV 결과이다. 가로축은 벌점모수에 자연로그를 취한 값이고, 세로축은 이탈도다. 수직으로 있는 두 개의 점선은 1-표준편차 범위를 나타낸다. 10-fold CV 결과, 이탈도 변화를 최소로 만들어주는 λ(벌점 모수) 값은 0.012였고, 1-표준편차 범위에서의 최소값은 0.026이었다. Occam's razor 법칙을 이용하여 1-표준편차 최소값으로 벌점회귀모형을 구축하였다(Hastie et al., 2009).



[그림 2] 이탈도 기준 10-fold CV 결과

2. LASSO 결과

LASSO 후 회귀계수가 0이 아닌 설명변수는 15개였다(표 1 참고). 모든 변수가 표준화된 후 모형화되었으므로 이 계수들은 표준화 계수라고 할 수 있고, 그 상대적인 중요도 또한 도출할

수 있다. 참고로 벌점회귀모형에서는 회귀계수의 표준오차를 산출하지 않는다. 벌점회귀모형은 추정치의 분산을 줄이는 대신 편향이 늘어나는 방법이므로 평균제곱오차에서 편향이 큰 부분을 차지한다. 그러나 벌점회귀모형에서 정확한 편향을 측정하기 힘들기 때문에 표준오차나 신뢰구간을 구하는 것이 의미가 없기 때문이다(Goeman, Meijer, & Chaturvedi, 2016).

15개의 변수 중 중학교 졸업 후(향후 예상) 진로(q2k1w5)가 청소년의 진로선택 여부와 가장 큰 관련이 있었다. 중학교 졸업 후 고등학교 진학, 기타 진학, 또는 취업을 결정했다고 답할수록 진로선택 확률이 높았다. 다음으로 적성 및 소질에 대하여 잘 알지 못하거나(q1a1w5), 직업에 대한 정보 부족으로 직업의 종류, 성격에 대해 알지 못하거나(q1a2w5), 미래는 불확실하므로 직업을 미리 선택하는 것은 무의미하다고 생각할수록(q1a6w5) 진로미선택 확률이 높았다. 이 세 개의 직업선택 관련 문항들은 기존의 진로선택을 연구한 논문에서 반응변수로 주로 이용된 문항들로, 문항묶음을 통하여 빈번하게 사용되었다. 본 연구에서는 전체 7개 문항 중 3개 문항만이 선택되었다.

‘향후 진로 설정과 관련된 전반적인 의견’은 원래 18개 문항으로 구성되었는데, LASSO를 이용한 벌점회귀모형으로는 18개 문항 중 단지 3개 문항만이 선택되었다. 자기개발을 위해 상급학교 진학이 필수적이라고 생각하거나(q2m01w5) 사회생활을 일찍 경험하고 싶을수록(q2m07w5) 진로선택 확률이 높았고, 자신의 적성 및 소질이 무엇인지 아직 잘 알지 못한다고 생각할수록(q2m12w5) 진로미선택 확률이 높았다.

KYPS에서는 학생 및 부모에게 희망 교육 수준을 질문하였는데, 학생 본인이 희망하는 교육수준이 높을수록(q3w5) 진로선택 확률이 높았고, 부모가 희망하는 교육 수준은 진로선택 확률과 관련되어 있지 않았다. ‘지난 1년간 진로관련 활동’은 7개의 문항으로 구성되어 진로관련 연구에서 이용된 바 있다. 본 연구에서는 7개 중 2개 문항이 진로선택과 관련 있었다. 적성검사를 수행할수록(q6_3a3w5), 그리고 진로 관련 책/잡지를 열독할수록(q6_3a7w5) 진로선택 확률이 높았다.

지난 1년간 교내 경시대회 수상경력(q6l1w5)이 있을수록, 장래에 선생님과 같은 사람이 되고 싶다고 답할수록(q9a07w5), 자신의 삶을 스스로 책임지며 살고 있다고 답할수록(q21a3w5) 진로선택과 확률이 높았다. 또한 심리적 혹은 정신적으로 문제가 있을수록(q2i02w5), 그리고 현재 다니는 학교가 특기나 소질을 살리는데 한계가 있을수록(q2i11w5) 진로미선택보다 진로선택 확률이 높았다.

<표 1> LASSO 결과 0이 아닌 15개 설명변수의 회귀계수와 코딩 방식

	변수 설명	코딩	계수
q1a1w5	직업선택_적성 및 소질에 대해서 알지 못함	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	−0.333
q1a2w5	직업선택_직업에 대한 정보 부족으로 직업의 종류, 성격에 대해 알지 못함	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	−0.057
q1a6w5	직업선택_미래는 불확실하므로 직업을 미리 선택하는 것은 무의미한 일	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	−0.088
q2i02w5	현재 전반적인 상황_심리적 혹은 정신적으로 문제가 있다	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.017
q2i11w5	현재 전반적인 상황_현재 다니는 학교는 특기나 소질을 살리는데 한계가 있다	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.002
q2k1w5	중학교 졸업 후(향후 예상) 진로	0(아직 구체적으로 정하지 못함), 1(나머지)	0.483
q2m01w5	향후진로관련 의견_자기개발을 위해 상급학교 진학이 필수적	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.056
q2m07w5	향후진로관련 의견_사회생활을 일찍 경험하고 싶음	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.017
q2m12w5	향후진로관련 의견_나의 적성 및 소질이 무엇인지 아직 잘 알지 못함	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	−0.139
q3w5	희망 교육 수준(본인)	0(잘 모르겠다), 1(중졸), 2(고졸), 3(초대졸; 2~3년제), 4(대졸; 4년제), 5(대학원졸; 석사 및 박사)	0.013
q6l1w5	지난 1년간 교내 경시대회 수상경력	1(있다), 0(없다)	0.122
q6_3a3w5	지난 1년간 진로관련활동_적성검사_수행여부	1(있다), 0(없다)	0.036
q6_3a7w5	지난 1년간 진로관련활동_관련 책/잡지 열독_수행여부	1(있다), 0(없다)	0.043
q9a07w5	학교생활_나는 장래에 선생님과 같은 사람이 되고 싶다	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.010
q21a3w5	나는 내 삶을 스스로 책임지며 살고 있다	Likert: 1(전혀 그렇지 않다)~5(매우 그렇다)	0.039

* 변수는 KYPS 자료 입력 순으로 정리되었다.

Ⅴ. 논의

1. 연구 결과 논의

진로선택과 관련한 기존 KYPS 연구의 설명변수는 부모애착, 친구애착, 교사애착, 자아존중감, 자기효능감(자기신뢰감), 성별, 부모학력, 진로관련활동, 직업가치관 등이었다. 특히 부모애착은 9건의 연구 중 5건이 진로선택 설명변수로 이용한 인기 있는 변수였고, 자기효능감, 자아존중감, 진로체험활동도 두세 건의 연구가 모형에 이용했던 변수들이었다. 데이터 마이닝을 이용한 본 연구 결과, 부모 관련 변수와 친구 관련 변수는 그 계수가 모두 0으로 축소추정되었다. 성별과 직업가치관 또한 모형에서 선택되지 못했다.

반면, 김민선과 서영석(2010), 신선아와 전종설(2015)이 모형에 이용한 교사애착 3문항 중 한 문항인 '나는 장래에 선생님과 같은 사람이 되고 싶다'와, 김재철 외(2011)와 신선아와 전종설(2015)이 자기신뢰감 3문항 중 한 문항으로 이용한 '나는 내 삶을 스스로 책임지며 살고 있다'는 모형에 포함되었다. 허균(2010)이 모형화한 진로관련활동 6문항 중 적성검사 수행여부 및 관련 책/잡지 열독 여부, 그리고 김민선과 서영석(2010)의 '학교장벽' 중 한 문항인 '현재 다니는 학교는 특기나 소질을 살리는데 한계가 있다' 또한 모형에 포함되었다. 그 외 모형에서 선택된 '적성 및 소질에 대해 알지 못함' 등의 직업선택 관련 3문항은 선행 연구에서 진로성숙, 진로미결정 등의 반응변수로 이용된 문항이었다.

교사애착과 자기신뢰감이 높을수록 진로를 결정할 확률이 높았는데, 3개 문항 중 각각 하나의 문항만이 모형에 포함된 것이 특기할 사항이다. 진로관련활동 6개 문항 중 선택된 2개 문항을 고려할 때, 적성검사를 자주 수행할수록, 그리고 적성 관련 책/잡지를 열심히 읽을수록 진로 결정 확률이 높았다. 따라서 학교 또는 가정에서 적성검사를 보게 하고, 적성과 관련된 책/잡지 구독을 장려할 필요가 있는 것으로 보인다.

본 연구가 새롭게 탐색한 변수는 중학교 졸업 후 향후 예상 진로, 자기개발을 위한 상급학교 진학 의견, 본인의 희망교육 수준, 그리고 지난 1년간 교내 경시대회 수상경력 등이었다. 모두 중학생 대상 KYPS 선행 연구에서 다루지 않았던 것들로, 진로선택 여부와 정적인 관계가 있었다. 즉, 중학교 졸업 후 고등학교(예: 일반고, 특목고, 전문고 등) 진학 또는 취업을 결정했다고 답할수록 진로선택 확률이 높았다. 따라서 중학생을 대상으로 한 진로교육에서는 당장 앞으로 직면한 고등학교 진학 여부부터 초점을 맞출 필요가 있을 것으로 보인다. 자기개발을 위해 상급학교 진학이 필요하다는 인식을 심어주며, 학생들이 희망교육 수준을 어느 정도 높게 잡도록 진학지도를 하는 것도 학생들의 진로선택에 도움을 줄 것으로 생각된다. 교내 경시대회 또한 입상자들의 진로 탐색에는 도움이 되었던 것으로 보인다.

본인이 인식하는 심리적 혹은 정신적 문제 정도도 본 연구에서 새롭게 탐색된 변수로, 심리적 혹은 정신적으로 문제가 있다고 인식할수록 진로선택 확률이 높았다. 진로상담의 목적이 내담자의 심리적 문제를 극복하고 자신에게 적합한 진로를 선택하도록 하는 것(Gati, Krausz, & Osipow, 1996)임을 상기할 때 이는 해석하기 쉽지 않은 결과다. 이 문항을 모형화한 기존 KYPS 연구는 찾기 힘들었으나, 대학생의 진로스트레스에 대한 질적 연구에서 과하지 않은 심리적 문제는 자기통제 및 동기부여 역할을 하는 것으로 알려졌다(박미진 외, 2009). 본 연구의 KYPS 자료 분석 결과, 1부터 5까지 가능한 이 문항의 중앙값이 1, 평균값이 1.5에 불과하였으므로 본 연구의 중학생들이 과한 심리적 문제를 호소한 것이라고는 볼 수 없었다. 고등학생의 강박증 및 우울·불안과 진로미결정 관계를 연구한 박정희, 이은희(2008)의 위계적 회귀모형에서 강박증이 낮을수록, 그리고 우울·불안이 높을수록 진로결정 수준이 높았다는 결과를 찾을 수 있었다. 본 연구의 중학생들은 '심리적 혹은 정신적 문제'를 우울·불안증으로 받아들인 것이 아닌지 연구가설을 세울 수 있다. 심리적, 정신적 문제 유형(예: 우울·불안증, 강박증)에 따라, 그리고 심리적, 정신적 문제 정도에 따라 진로선택 확률이 어떻게 달라지는 것인지 후속 연구가 필요하다.

'현재 다니는 학교는 특기나 소질을 살리는데 한계가 있다'고 답할수록 진로선택 확률이 높았다. 이 문항은 김민선과 서영석(2010)의 연구에서 '학교장벽' 문항으로 이용된 적 있으나, 계수의 부호가 반대였다. 학생의 심리적, 정신적 문제와 달리, 이 문항 또는 이 문항과 비슷한 내용을 진로선택과 연결 지은 연구는 찾기 힘들었다. 본 연구 결과 계수의 절대값이 가장 작은 0.002로, 거의 0에 가까운 값이었으므로 학생의 진로선택에 상대적으로 가장 적은 영향을 미친 문항이라고 볼 수 있다. 진로와 관련하여 지대한 관심이 있는 학생들일수록 그 높은 기대치로 인하여 재학 중인 학교의 진로교육에 대한 만족도가 오히려 떨어지는 것이 아닌지 연구가설을 세울 수 있다. 역시 이를 검정할 수 있는 후속 연구가 필요하다.

2. 연구 함의

Ridge, LASSO, Elastic Net 등의 벌점회귀모형은 계수를 축소추정하는 특징으로 인하여 빅데이터 분석 시 필수적인 방법 중 하나로 여겨진다. 본 연구는 최근 각광받고 있는 벌점회귀모형 중 하나인 LASSO를 데이터 마이닝 기법으로 KYPS 자료에 적용함으로써 교육 패널자료에 대한 데이터 마이닝 기법의 적용 실제를 제시하였다. 즉, 300개가 넘는 변수를 한 모형에 모두 투입함으로써 선행연구에서 전혀 다루지 않았던 변수, 선행연구에서 한 모형에 투입되지 않은 변수 등을 새롭게 발굴할 수 있었다.

기존 연구에서 주로 쓰인 HLM, SEM, ARCL, LGM 등의 일반 회귀모형으로는 몇 천 명의 응답자가 있다고 하더라도 수백 개의 변수를 모두 한 모형에서 연구하기 힘들다. 수렴

(convergence) 또는 과적합(overfitting) 등의 문제가 발생할 수 있기 때문이다. 그러나 LASSO와 같은 데이터 마이닝 기법으로는 수백 개의 변수를 한 모형에서 연구하는 것이 가능하다. 벌점회귀모형은 표본 수보다 변수 수가 많은 자료에도 문제없이 이용될 수 있기 때문이다. 다시 말해, 본 연구는 빅데이터 분석 시 이용되는 데이터 마이닝 기법을 적용함으로써 선행연구와 차별되는 결과를 도출할 수 있었다. 이는 많은 비용과 노력을 들여 수집한 종단연구 자료를 제대로 활용하는 것뿐만 아니라, 그 결과를 통해 관련 내용 연구자들에게 새로운 시각을 제공해 줄 수 있다는 데 의의가 있다.

KYPS와 같은 교육 패널자료는 다양한 영역을 수백 개의 문항으로 측정하며 리커트 척도뿐만 아니라 범주형으로 자료를 수집하는 것이 특징이다. KYPS의 범주형 변수들은 문항들이 매우 자세하게 범주화되어 상세한 자료를 얻을 수 있는 점은 장점이지만, 동시에 자료 빈도의 불균형 문제 및 무응답이 많은 점은 모형화 시에는 단점이 될 수 있다. KYPS의 범주형 변수를 모형화하기가 쉽지 않았으므로 선행연구들은 주로 리커트 척도로 측정된 문항들을 묶어서 재구성하여 변수로 이용하였다. 특히 SEM, LGM 등과 같은 방법론으로 분석하는 연구에서 리커트 척도로 측정된 문항에 대한 문항묶음(item parcel)은 필수적이다. 선행연구 장에서 분석된 바와 같이 KYPS 설문지에서 같이 묶인 문항세트 전체(예: '직업선택' 항목의 7개 문항)가 주로 이용된 반면, 연구에 따라 문항세트의 일부만을 이용하거나 다른 잠재변수로 묶는 경우도 있었다(예: 김민선, 서영석(2010), 신선아, 전종설(2015) 등 참고). 그러나 이렇게 서로 다른 문항묶음을 같은 이름으로 명명하는 경우 해석에 매우 주의하여야 한다. 같은 이름의 변수라도 다른 문항묶음으로 측정되었기 때문이다.

또 다른 문제로, SEM, LGM 등과 같은 방법론에서 리커트 척도를 동간척도인 것처럼 취급하여 문항묶음으로 그 합이나 평균을 이용하여 하나의 변수를 만드는 것이 있다. 그러나 리커트 척도로 측정된 문항은 서열척도이므로 동간척도처럼 이용하는 것에는 무리가 따른다. 이는 문항묶음으로도 해소되지 않는 문제라 할 수 있다. 더욱 근본적인 문제는 하나의 문항묶음을 구성하는 하위 문항들의 내적합치도가 높아야 한다는 점이다. 그런데 그 값이 매우 낮은 경우가 빈번하였다. 이를테면 허성호와 정태연(2010)의 '자기신뢰' 척도는 크론바흐 알파 계수가 0.32에 불과하였다. 그 외 다른 연구들도 크론바흐 알파 계수가 0.8 미만인 경우가 흔하였다. 크론바흐 알파 계수가 적어도 0.8 이상은 되어야 한다는 것이 문항묶음의 일반적인 원칙임을 감안할 때, 문항묶음을 이용할 필요가 없는 벌점회귀모형이 대안 중 하나가 될 수 있을 것으로 보인다.

정리하자면, 데이터 마이닝 기법 중 벌점회귀모형을 이용한 본 연구의 장점으로 수백 개가 넘는 변수들을 모두 한 모형에 투입하여 중요한 변수를 선택할 수 있는 점, 그리고 리커트 척도로 측정된 문항 분석 시 문항묶음을 고심할 필요가 없는 점 등을 들 수 있다. 주의할 점으로, 데이터 마이닝 기법의 특성 상 연구의 이론적 배경보다는 통계적 자료 분석을 통한 변수 추출이

주가 되므로 기존 연구의 이론적 배경을 중시하는 접근법과는 근본적으로 다른 관점에서 분석이 진행된다. 본 연구에서도 이러한 데이터 마이닝 기법의 특성을 견지하며 분석을 수행하였으며, 그 결과를 선행연구 결과와 비교·분석하여 제시하였다.

참고문헌

김민선, 서영석(2010). 자기효능감, 개인배경, 맥락적 변인이 청소년의 진로미결정 수준 변화에 미치는 영향에 관한 종단연구. **한국청소년연구, 21**(2), 67-96.

김보현, 하일도, 노맹석, 나명환, 송호천, 김자혜(2015). frailtyHL 통계패키지를 이용한 프레일티 모형의 변수선택: 유방암 생존자료. **응용통계연구, 28,** 965-976.

김재철, 황매향, 김아영(2011). 체험활동과 진로성숙 간의 관계에서 긍정적 자아관과 내적 직업가치관의 매개효과. **진로교육연구, 24,** 1-23.

박미진, 김진희, 정민선(2009). 진로상담 : 취업준비 대학생의 스트레스에 대한 질적 연구. **상담학연구, 10,** 417-435

박민수, 김태헌, 조은석, 김희발, 오희석(2014). R을 이용한 벌점화 축소추정 기법 비교연구: 요크셔 돼지 산자수와 SNP에 대한 적용 사례. **농업생명과학연구, 48,** 147-155.

박정희, 이은희(2008). 청소년의 자아 정체성, 불안/우울 및 강박증과 진로미결정: 자기 통제력과 사회 지원의 매개역할. **한국심리학회지: 상담 및 심리치료, 20,** 103-123.

박철용, 계묘진(2013). 호흡곤란 환자 퇴원 결정을 위한 벌점 로지스틱 회귀모형. **한국데이터정보과학회지, 24,** 125-133

송상윤(2015). 예대금리차 결정요인 모형의 예측력 비교 연구: Ridge, LASSO 및 Elastic Net 방법론을 중심으로. **금융지식연구, 13,** 41-65.

신선아, 전종설(2015). 청소년의 애착이 자기효능감을 매개로 진로성숙도에 미치는 영향. **청소년복지연구, 17**(3), 111-136.

어윤경(2008). 진로교육 만족도에 따른 진로성숙 수준 변화에 대한 다층분석. **진로교육연구 21**(4), 23-41.

진슬기, 김광래, 박창이(2012). 신용평점화에서 벌점화를 이용한 절단값 선택. **응용통계연구, 25,** 261-267.

허균(2010). Autoregressive Crosslagged Model을 적용한 진로경험활동과 진로성숙도의 종단관계연구. **직업교육연구, 29,** 157-170

허균(2012a). 자아존중감과 진로장벽의 자기회귀교차지연 효과 분석 연구. **직업교육연구, 31,** 119-134.

허균(2012b). 잠재성장모형을 활용한 진로성숙도의 변화궤적과 성별, 자아존중감 및 부모애착 시간효과의 구조관계. **직업교육연구, 31,** 193-209.

허균(2013). 초기청소년의 발달 과정에서 진로성숙도에 대한 부모애착의 동시효과와 지연효과

연구. **직업교육연구, 32,** 107-118.

허성호, 정태연(2010). 자원봉사활동이 청소년기 발달에 미치는 영향. **한국청소년연구, 21**(3), 143-164.

Fan, H., Han, F., & Han, L. (2014). Challenges of big data analysis. *National Science Review, 1,* 293-314.

Gati, I., Krausz, M., & Osipow, S. H. (1996). A taxonomy of difficulties in career decision making. *Journal of Counseling Psychology, 43,* 510-526.

Goeman, J., Meijer, R., & Chaturvedi, N. (2016). *L1 and L2 penalized regression models.* Retrieved May 30, 2016 from https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Korea Youth Panel Survey. (n. d.). *Korea Youth Panel Survey.* Retrieved May 30, 2016 from http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=316&sitesectiontitle=Korea+Youth+Panel+Survey.

Lange, K., Papp, J. C., Sinsheimer, J. S., & Sobel, E. M. (2014). Next generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. *Annual Review Statistical Application, 1,* 279-300.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistics Society B, 58,* 267-288.

Varian, H. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28,* 3-28.

* 논문접수 2016년 8월 2일 / 1차 심사 2016년 9월 12일 / 게재승인 2016년 9월 21일

* 유진은: 서울대학교 교육학과 졸업(B. A.) 후 미국 Purdue University에서 교육학 석사(M. S. in Education), 통계학 석사(M. S. in Statistics) 취득 후 측정 · 평가 · 연구방법론으로 박사학위(Ph. D.)를 받았다. 주요 저서로는 '한 학기에 끝내는 양적연구방법과 통계분석', 'Multiple Imputation with Structural Equation Modeling' 등이 있다.
* E-mail: jeyoo@knue.ac.kr

Abstract

An Analysis Case of Educational Panel Data through a Data Mining Technique: A Penalized Regression with KYPS Data

Yoo, Jin Eun*

With the advent of so-called big data era, data mining techniques have come to the fore as big data analysis tools. Unlike conventional statistical methods, data mining techniques can handle hundreds of variables in one model without convergence or overfitting problems. However, studies in the field of education have not yet paid enough attention to recent data mining techniques. Particularly, panel data with its hundreds of variables and thousands of participants can fit data mining techniques. This study aimed to illustrate a popular data mining technique, LASSO, by applying it to the 5th wave of KYPS (Korea Youth Panel Study). A penalized LASSO regression was executed with 10-fold cross-validation via deviance, and was successfully applied to the social sciences panel data. Implications of the study are discussed as well as further research topics.

Key words: educational panel data, data mining, big data, penalized regression, KYPS

* Associate Professor, Korea National University of Education