The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.7.1125
27(7), 1125–1137

The Analysis of Private Education Cost for the Elementary, Middle, and High School Students in Korea

Hyejeong Lee[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received September 16, 2014; Revised October 14, 2014; Accepted November 06, 2014)

Abstract

This paper studies what affects the private education cost for the elementary, middle, and high school students. It is a big issue now because there can be a problem in the equal opportunity for education if the portion of private education cost is very high in the total education cost. If we spend more time and money on the private education than the school education, it can cause the polarization among the classes and regions. The excessive private education also can deteriorate the school system. we use various regression and classification methods to analyze the cost of private education and find the important variables in the models. we found that large cities spend more money on the private education than small cities. We also found that high school students spend more than middle school students and the elementary students and the household with more income spend more money on the private education.

Keywords: Private education cost, regression, classification, random forests, gradient boosting

1. 서론

현재 한 포털 사이트 검색란에 "사교육비"를 검색하면 1년 사이에 7천여 건의 기사가 검색되는 만큼 현 사회에서의 사교육에 대한 관심과 문제는 상당하다. 이러므로 현재 정부에서는 사교육비에 대하여 파악하고자 2007년부터 우리나라의 초, 중, 고 학생들의 사교육 실태를 조사하였다. 과거 6 년간 통계청에서 발표된 사교육비의 월 평균은 Figure 1.1과 같다. 2007 년 처음 조사했을 당시 22만 2천 원에서 2009년 24만 2천 원으로 월 평균 사교육비가 3년간 계속 증가하였고(통계청 사교육비 조사), 최근에 조사된 2012년의 월 평균 사교육비는 23 만 6 천 원이다(통계청, 2013). 2008년 2009년의 경제위기와 사교육 포화 상태, 정부의 개입 등으로 인하여 더 이상의 월 평균 사교육비는 증가하지 않았지만, 물가 상승에 따른 사교육비에 대한 부담은 늘어가는 실정이다. 이러한 관심 속에 사교육비 자료를 이용한 선행 연구가 많이 행해졌는데, 많은 연구 가운데 우리는 2008년 사교육비 자료를 이용한(김진희, 2010) 연구

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the ministry of Education, Science and Technology (No. NRF-2013R1A1A2012817)

[1]Corresponding author: Associate Professor, Department of Statistics, Ewha Womans University, Seoul, 120-750, Korea. E-mail: josong@ewha.ac.kr

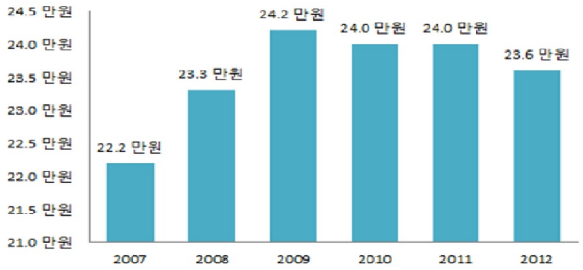

Figure 1.1. 2007 ∼ 2012 Monthly average private education cost

와 2009년 사교육비 자료를 이용한 새로운 사회를 여는 연구원(김수현, 2010) 연구, 2012년 통계청 조사 결과 보도 자료를 참고하여 과거의 사교육비 경향을 살펴본 뒤 2012년 사교육비 자료를 이용하여 분석을 하였다.

2. 방법론

본 논문에서는 회귀분석과 분류분석 방법론을 이용하여 자료를 분석하였다. 회귀분석에서는 선형회귀모형(linear regression model)과 stepwise regression(Venables와 Ripley, 2003), random forests(Breiman, 2001), gradient boosting(Friedman, 1999; Ridgeway, 2012) 방법론을 사용하였고 분류분석에서는 logistic regression, random forests, gradient boosting 방법론을 사용하였다. 이 절에는 이런 방법론을 간단히 설명한다.

2.1. 선형회귀모형(linear regression model)

선형회귀모형은 반응변수가 설명변수의 선형식으로 설명될 수 있다고 가정하고 모형을 적합하는 방법론이다. 만약 관측치의 개수가 n이고, 설명변수가 p개가 있다고 가정하면 일반적으로 다음과 같이 모형을 표현할 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_i p + \epsilon, i = 1, \cdots, n.$$

ϵ는 오차 항으로 평균이 0, 분산이 σ^2 라고 가정한다. 선형회귀모형을 적합하기 전에 all pairwise plot을 이용하여 반응변수와 설명변수들 간에 정말 선형관계가 있는 지를 확인하는 절차는 꼭 필요하며 모형 적합이 끝난 후에는 잔차 산점도를 확인하여 오차항에 대한 가정이 맞는 지를 점검하는 것도 필수적인 절차이다. 모수추정은 잔차제곱합을 최소화하는 최소제곱추정량(least squared estimator) 방법을 이용한다. 만약 설명변수의 수가 많은 경우에는 변수선택방법을 이용하여 모형에 유용한 변수를 선택하는데 일반적으로 stepwise regression 방법론을 많이 이용한다. R에서 구현된 stepwise regression 함수는 AIC값을 최소화하는 모형을 매 스텝마다 설명변수를 하나씩 모형에 추가하거나 제거하면서 찾는다. 물론 이 방법이 AIC를 가장 최소화하는 모형을 찾는 것을 보장하지는 못하지만 그래도 적은 계산과정으로 경쟁력 있는 모형을 찾을 수 있으므로 많이 선호되는 방법론이다(박창이 외, 2011).

Table 3.1. Distribution of students

	초등학생	중학생	고등생(일반)	고등생(전문)	총합
빈도	20,605명	22,239명	35,706명	7,632명	86,182명
비율	23.9%	25.8%	41.4%	8.86%	100%

2.2. Random Forests

Breiman이 2001년도에 제안한 random forests 방법론은 tree bagging 방법론으로 일반적인 tree bagging 방법론과 다른 점은 각각의 bootstrap sample에서 적합된 tree들의 상관계수를 낮춤으로서 전체적인 성능을 증가시키는 데 있다(Hastie 등, 2009). 이는 tree를 만들어 나갈 때 splitting variable을 전체 설명변수에서 선택하지 않고 랜덤한 작은 그룹에서 선택함으로써 가능해진다. Random forests는 회귀분석과 분류분석 모두에서 사용할 수 있으며 높은 예측 정확성 때문에 많이 사용되고 있다. 또한 random forests는 variable importance measure 를 제공함으로써 어떤 변수가 예측에서 중요변수인지를 알려주며 설명변수의 수가 관측치의 수보다 크더라도 문제없이 적합할 수 있다.

2.3. Gradient Boosting

Graident boosting은 1999년에 Friedman 이 제안했으며 추후 개선된 방법론이 제안되었다. Gradient boosting은 손실함수를 최소화는 함수를 추정하는 것이 목표이며 일반적으로 작은 크기의 tree를 사용하고 때로는 1개의 splitting variable만 있는 stump를 사용하는 경우도 있다. Random forests와 같이 test set에서 높은 예측 정확성을 보여주며 회귀분석과 분류분석 모두에서 사용가능 하지만 multi-class 분류에서는 아직 사용이 용이하지 않다.

3. Data

이번 절에서는 본 논문에 사용된 자료에 대해서 설명할 것이다.

3.1. 조사 대상 및 기간

본 연구에서 사용한 자료는 2012년 우리나라 초, 중, 고 학생들의 사교육비 조사 자료이다. 본 자료는 총 86,182 개의 관측치와 53개의 변수로 이루어져 있다. 총 86,182명의 초, 중, 고 학생 분포는 Table 3.1과 같다.

2012년 사교육비는 1년에 2번 조사되었다. 3, 4, 5월분의 사교육비를 6월 1일부터 20일 까지 1 차로 조사하였으며, 2 차 조사는 7, 8, 9월분의 사교육비를 9월 26일부터 10월 17일 까지 조사하였다.

3.2. 조사 방법

통계청에서 주관하는 사교육비 조사는 3가지 방법으로 실시하였다. 3가지 방법으로는 인터넷 설문 조사와 설문지 배포를 통한 방법 그리고 2012년에 처음 실시된 모바일을 통한 조사 방법이다. 혼합 조사로 인터넷 조사를 기본으로 종이 조사를 보충적으로 실시하고, 배포 조사로는 표본 학급의 담임교사가 학생에게 설문지를 전달하여 학부모가 작성한 뒤 회수 하는 방법을 이용하였다.

Table 3.2. Survey items for 2012 private education cost

항목	변수 명	
기본조사문항	지역 구분	행정구역 (시도)
	학교 급 구분	일반교과 사교육 목적 구분 1
	일반교과 사교육 목적 구분 2	일반교과 사교육시간
	예체능 사교육 목적 1	예체능 사교육 목적 2
	예체능 사교육시간	취업관련 사교육시간
	사교육 참여 시간	학생성별
	학생 성적 구분	부의 연령
	부의 교육정도	모의 연령
	모의 교육정도	부모의 경제활동
	가구 소득	
방과 후 학교 문항	방과 후 학교 참여(유상+무상)	방과 후 학교 참여시간
	방과 후 학교 (금액)	초등방과 후 보육프로그램(금액)
	특기적성 프로그램 (금액)	교과 프로그램 (금액)
기타 문항(금액)	EBS 교재비	어학연수비
	국내연수	해외연수
사교육비(금액)	사교육비	
일반교과 사교육 문항(금액)	일반교과 사교육비	국어
	영어	수학
	사회, 과학	제2외국어, 한문, 컴퓨터
	논술	일반교과 개인과외
	일반교과 그룹과외	일반교과 학원수강
	일반교과 방문 학습지	일반교과 유료인터넷 및 통신강좌
예체능 교과 사교육 문항(금액)	예체능, 취미, 교양 사교육비	음악
	미술	체육
	취미, 교양	예체능 개인과외
	예체능 그룹과외	예체능 학원수강
	예체능 방문수업, 유료인터넷 및 통신강좌	
취업문항	취업관련 (고등학교의 경우)	

3.3. 조사 변수

조사는 크게 학생에 대한 기본 조사 문항과 방과 후 학교에 관한 문항, 일반교과 사교육에 대한 문항, 예체능 사교육에 대한 문항, 취업 사교육에 관련한 문항으로 이루어져있다. 세부 조사된 항목은 Table 3.2와 같다.

3.4. Data preparation

A. 결측값 처리 조사된 변수들 가운데 본 연구와 관계가 적고, 0원의 비율이 높은 취업 관련 변수와 기타 문항의 EBS 교재비, 어학연수비, 국내연수, 해외연수 변수는 분석에서 제외하였다. 또한, 52개의 변수 가운데 아버지의 연령, 아버지의 교육정도, 어머니의 연령, 어머니의 교육정도의 4개의 변수에서 결측값이 관측되었다. 총 86,182의 응답 가운데 결측값은 7,652개로 10% 미만으로 적고, 정확한 분석을 위하여 삭제하기로 한다. B. 단위의 표준화 시간에 관련한 변수는 주 단위이고, 금액에 관련한 변수는 연 단위이다. 따라서 시간에 관련한 변수에는 4배를 하여 월 단위로 바꾸었고, 연 단위의 변수는 12로 나누어 월단위로 바꾸어 분석하였다. C. 변수형 정의 변수형은 총 2가지로 동일한 성질을 가진 부류나

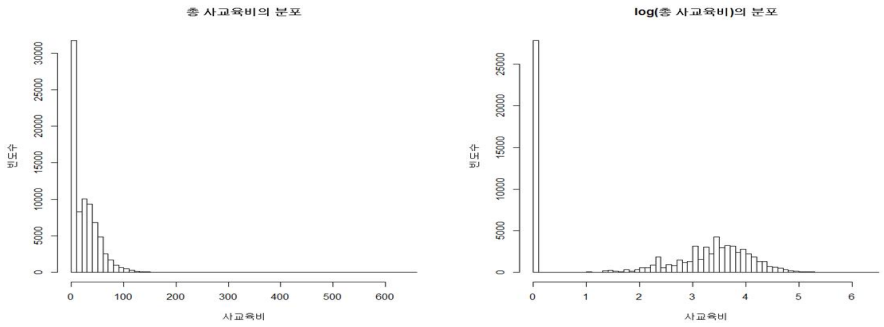

Figure 3.1. Histogram of total private education cost and the histogram of the total private education cost after log-transformed

Table 3.3. Distribution of total private education cost

	최소값	1사분위	중앙값	평균값	3사분위	최대값
변환 전	0.00	0.00	17.8	23.61	38.2	652.8
log변환 후	0.00	0.00	2.93	2.19	3.67	6.48

범위로 이루어진 범주형과, 숫자 값으로 된 연속형이다. 변수들 가운데 범주형 이지만 숫자가 증가함에 따라서 자명하게 사교육비가 증가하는 변수들은 연속형으로 간주 하였으며, 각 구분마다 특징이 있는 변수들은 범주형으로 변환 하여 dummy변수로 간주 하였다. 범주형으로 변환한 변수들은 지역 구분, 학교 급 구분, 학생 성별, 부모의 경제 활동, 방과 후 학교 참여 총 5개의 변수이며, 범주형인데 연속형으로 분석한 변수는 지역 구분, 학생 성적 구분, 아버지 연령, 아버지 교육정도, 어머니의 연령, 어머니의 교육정도, 가구 소득이다. D. 이상값 처리 총 사교육비를 보면 응답 가운데 일반교과 사교육시간과 예체능교과 사교육시간은 0인데 반해 총 사교육비가 0이 아닌 응답이 있다. 총 317개로 이상값이라고 생각 하여 제외 하였다.

3.5. 사교육비에 분포

A. 총 사교육비 총 사교육비는 일반교과와 예체능교과의 사교육비를 더한 사교육비를 뜻한다. 이때 사교육은 월 단위 이고, 만원 단위이다. Figure 3.1 히스토그램을 보면 사교육비가 0인 자료가 많은 반면 몇 개의 너무 큰 값이 있어서 히스토그램이 오른쪽으로 긴 꼬리를 가지고 있다. 따라서 log 변환을 실시하였다. 하지만 0인 자료에 대해서는 -INF 값이 나오기 때문에 모든 사교육비 값에 +1을 하여서 log 변환을 실시하였다. 아래의 Table 3.3에서 볼 수 있듯이 25% 이상의 학생들이 사교육비가 0인 것을 볼 수 있고, 평균은 23.61만원, 최댓값은 약 653만원 이었다.

B. 일반교과 사교육비 일반교과 사교육비는 총 사교육비와 마찬가지로 월 단위의 사교육비 이며, 만원 단위 이다. 여기에 포함되는 과목은 국어, 영어, 수학, 사회, 과학, 제2외국어, 한문, 컴퓨터, 논술에 해당한다. Table 3.4와 같이 일반교과 사교육비도 마찬가지로 왼쪽으로 치우친 모양의 분포를 보이기 때문에 log 변환을 위해 모든 일반교과 사교육비에 +1 씩 하여 진행 하였다.

C. 예체능교과 사교육비 예체능교과 사교육비 예체능교과 사교육비 역시 월 단위, 만원 단위의 금액이다. 여기에 포함되는 과목은 음악, 미술, 체육, 취미, 교양 이며 다른 종속변수들 보다는 0원의 비율이 상당히 크다. 중앙값이 0인 사실로 보아 상당수의 학생이 예체능교과에 대해서는 사교육을 시행하지 않

Table 3.4. The distribution of the general class private education cost

	최소값	1사분위	중앙값	평균값	3사분위	최대값
변환 전	0.00	0.00	9.90	19.84	32.70	652.8
log변환 후	0.00	0.00	2.93	1.87	3.52	6.48

Table 3.5. The distribution of the arts and physical private education cost

	최소값	1사분위	중앙값	평균값	3사분위	최대값
변환 전	0.00	0.00	0.00	3.61	0.00	324.00
log변환 후	0.00	0.00	0.00	0.59	0.00	5.78

는 것을 알 수 있다. 이 예체능교과 사교육비 또한 +1 씩 하여서 log 변환을 실시한 뒤 분석한 결과는 Table 3.5와 같다.

4. 분석 결과

4.1. 분류 분석 결과

종속변수인 사교육비 응답에 0원의 응답이 전체 응답의 30%를 차지하고 있기 때문에 정확한 분석에 어려움이 있었다. 따라서 우리는 분류분석을 통하여 사교육비 0원과 0원 초과의 자료를 분류하는데 영향을 미치는 변수를 알아보고자 하였다. 즉 사교육을 시행하는 학생과 시행하지 않는 학생을 분류할 수 있는 변수가 무엇인지 알아보고자 하였다. 이때 사용한 도구로 변수의 중요도를 계산하였다. 변수 중요도 값은 퍼센트 값으로 logistic regression 경우에는 분산분석표를 구하여 각 변수에 해당하는 sum of square 값을 sum of square 값의 전체 합으로 나눠 얻은 값이며, random forests와 gradient boosting 분석 방법에서는 통계 프로그램 중 하나인 R 에서 제공하는 relative influence를 이용하여 얻은 결과 값으로 각 변수가 모형을 만드는데 차지하는 상대적인 설명력을 나타내었다. 분석에 앞서 사교육 시행 여부에 따라서 사교육비 응답이 0원이면 “0”, 사교육비 응답이 0원 초과면 “1”로 변환한 새로운 종속변수를 만들어 분석을 실시하였다. 그 결과 “0”이 27,807개, “1”이 50,406 개로 자료가 구성되어 있었다. 분류분석은 세 가지의 방법으로 실시하였는데 logistic regression과 random forests, gradient boosting 방법이다. 또한, 일반교과 및 예체능교과 사교육시간이 중요변수인 것은 너무나도 당연하며, 중요변수를 구하는데 있어서도 너무 영향력이 큰 변수이므로 다른 변수들에 대해서 더 알아보고자 사교육시간이 없는 경우의 중요 변수에 대해서도 추가로 분석하였다. 그 결과는 Table 4.1과 같으며, 이때 얻은 오분류율은 10-fold CV 계산을 100번 반복 시행해서 얻은 오분류율 결과의 평균이다. 일반교과 및 예체능교과 사교육시간이 포함된 경우의 분류 분석의 오분류율은 0%로 성능이 완벽하다. 하지만 일반 및 예체능교과 사교육시간을 제외한다면 각 총 사교육비와 일반교과 및 예체능교과 사교육비에 대해서 27%, 30%, 19% 의 오분류율을 얻는다. 분류 분석의 결과로 얻어진 사교육비 0원 여부에 따른 분류의 기준이 되는 변수는 Table 4.2과 같다. 이때 logistic regression의 경우 적합과정에서 경고가 발생하였고, 최종 모형의 유의성은 아주 낮지만 그럼에도 불구하고 예측력은 좋았다. 총 사교육비, 일반교과 및 예체능교과 사교육시간이 포함된 분석에서는 일반교과 사교육시간과 예체능교과 사교육시간, 학교 급 구분 등이 중요변수로 다수 선택되었고, 사교육시간이 제외된 분석에서는 학교 급 구분, 가구 소득, 학생 성적 등이 다수 선택되었다. 공통으로 선택된 변수 가운데 학교 급 구분에 대하여 살펴보면, 전문 고등학교일 때 0원에 대한 비율이 높아서 분류하는데 중요변수로 선택 되었을 것으로 생각된다. 또한, 가구 소득이 낮을 때 0원에 비율이 높고, 일반교과의 경우 성적이 낮을수록 0 원의 비율이 높고, 예체능교과의 경우에는 작은 차이지만 성적이 상위권일 때 0원의 비율이 높으므로 중요 변수로 선택되었다고 판단된다.

Table 4.1. Misclassification rate

오분류율		Logistic Regression	Random Forests	Gradient Boosting
사교육 시간포함	총 사교육	0%	0%	0%
	일반교과	0%	0%	0%
	예체능교과	0%	0%	0%
	총 사교육	27%	27%	27%
사교육 시간제외	일반교과	31%	30%	31%
	예체능교과	20%	19%	20%

Table 4.2. Important variables in the classification

		Logistic Regression		Random Forests		Gradient Boosting	
		중요변수	(%)	중요변수	(%)	중요변수	(%)
시간 포함	총	일반교과 사교육시간	58	일반교과 사교육시간	67	일반교과 사교육시간	81
		예체능교과 사교육시간	31	예체능교과 사교육시간	22	예체능교과 사교육시간	20
		학교 급 구분	9	학교 급 구분	4	-	-
	일반	일반교과 사교육시간	91	일반교과 사교육시간	90	일반교과 사교육시간	100
		학교 급 구분	7	학교 급 구분	3	-	-
		지역 구분	1	방과 후 학교 참여여부	2	-	-
	예체능	예체능교과 사교육시간	84	예체능교과 사교육시간	86	예체능교과 사교육시간	100
		학교 급 구분	16	학교 급 구분	7	-	-
		-	-	모의 연령	2	-	-
시간 제외	총	학교 급 구분	53	일학교 급 구분	18	학교 급 구분	48
		가구 소득	12	가구 소득	16	가구 소득	6
		학생 성적	9	방과 후 학교 참여시간	14	학생 성적	6
	일반	학교 급 구분	48	학교 급 구분	18	학교 급 구분	43
		학생 성적 구분	16	가구 소득	16	가구 소득	29
		가구 소득	14	방과 후 학교 참여시간	14	학생 성적	13
	예체능	학교 급 구분	87	학교 급 구분	30	학교 급 구분	84
		부의 교육정도	3	방과 후 학교 참여시간	11	방과 후 학교 참여시간	5
		방과 후 학교 참여시간	2	가구 소득	10	모의 교육정도	3

4.2. 회귀분석 결과

이 절에서는 stepwise regression을 이용한 linear regression model, random forests, 그리고 gradient boosting 방법론을 이용하여 자료를 분석하였다. 그리고 모든 경우에 사교육비는 log 변환을 한 후에 분석을 시행하였다.

4.2.1. 모형 최적화 먼저 자료에 대한 분석 전에 3개의 모형에 대한 자료에 맞는 모수와 모형을 찾기 위하여 자료를 training data (70%) 와 test data(30%) 로 나누어서 test data에서 mean squared error (MSE)가 가장 작아지는 tuning parameter들을 선택하였다. Linear regression 모형 분석은 R에서 제공하는 step 함수를 사용하여 AIC값을 최소화하는 최적의 모형을 찾고, random forests 모형은 R package random forests 함수에서 제공하는 기본 옵션값을 사용하여 분석 하였다. Gradient boosting 모형의 경우에는 여러 가지 tuning parameter들이 있으나 우리는 shrinkage parameter만을 고려하였다. 이 parameter값을 0에서 0.1 까지 0.01 단위씩 증가해 가면서 test set에서의 MSE를 계산한 결과 0.01에서 가장 작은 MSE이 나왔으므로 이 값을 분석에 사용하였다.

Table 4.3. The average of MSE in a test set

MSE	총 사교육비	일반 사교육비	예체능 사교육비
Linear Regression	0.35	0.33	0.38
Random Forests	0.28	0.28	0.33
Gradient Boosting	0.29	0.28	0.35

Table 4.4. The result of linear regression model for the total private education cost

	회귀계수	표준오차	t	유의확률
절편	2.2605	0.021	107.795	2.00E-16
지역: 광역시	-0.1333	0.0087	-15.296	2.00E-16
지역: 중소도시	-0.1541	0.0084	-18.28	2.00E-16
지역: 읍면지역	-0.2457	0.0109	-22.618	2.00E-16
학교: 중학교	0.2282	0.0075	30.234	2.00E-16
학교: 일반고	0.516	0.0082	62.828	2.00E-16
학교: 전문고	0.1417	0.0187	7.596	3.10E-14
일반교과 사교육시간	0.0167	0.0001	118.675	2.00E-16
예체능교과 사교육시간	0.0088	0.0002	46.783	2.00E-16
성적 구분	-0.0275	0.0022	-12.454	2.00E-16
부 교육정도	0.0447	0.0049	9.117	2.00E-16
모 연령	0.0273	0.0058	4.67	3.00E-06
모 교육정도	0.0583	0.0054	10.841	2.00E-16
경제활동: 어머니 외벌이	-0.0696	0.0132	-5.26	1.50E-07
경제활동: 맞벌이	-0.0466	0.0056	-8.364	2.00E-16
경제활동: 참여안함	-0.0928	0.0287	-3.237	1.20E-03
가구 소득	0.0734	0.0016	45.543	2.00E-16
방과 후 학습: 참여	-0.0574	0.0071	-8.079	6.70E-16
방과 후 학습 참여시간	-0.0023	0.0002	-10.856	2.00E-16

4.2.2. 모형 성능 비교 앞에서 구한 모형을 이용하여 3가지 회귀분석 방법론의 성능을 비교하였다. 우리는 training data 와 test data를 7:3 으로 랜덤하게 나누어 training data에서 모형을 적합하고 test data에서 MSE 를 계산하는 과정을 100번 시행하였고 그 결과는 Table 4.3에서 볼 수 있다. Table 4.3에서 볼 수 있듯이 random forests가 3가지 경우 모두에서 가장 좋은 성능을 보여주었고 gradient boosting도 거의 비슷한 성능을 보여주었다. linear regression model은 가장 떨어지는 성능을 보여주었지만 그 차이가 별로 크지 않고, 모형 설명력에 있어서는 가장 좋은 모형이므로 앞으로의 분석에도 계속 사용할 것이다. 앞으로 분석에는 모든 training과 test로 나누지 않은 모든 자료를 이용하여서 최종 결과를 내릴 것이며, linear regression은 모형에 대한 해석과 함께, random forests와 gradient boosting 방법과 같이 비교하여 각 변수들에 대한 모형에 영향을 끼치는 정도를 알아봄으로써 중요변수를 알아내는 해석을 중점으로 진행하겠다.

4.2.3. 회귀모형 분석 결과 교육시간을 분석에 포함한 경우와 포함하지 않은 경우의 회귀분석 결과는 Table 4.4와 같다.

A. 교육시간을 분석에 포함한 경우

총사교육비에 대해서 stepwise regression을 사용한 linear regression model의 결과는 다음과 같다.

먼저 지역에 따른 월 평균 총 사교육비의 특징을 살펴보자. 다른 모든 조건들이 동등하고 지역만 다를

Table 4.5. Important variables in the linear regression model for the total education cost

변수	변수 중요도(%)
일반교과 사교육시간	47.29
학교 급 구분	20.82
지역 구분	10.31
예체능교과 사교육시간	7.33
가구 소득	6.51
부의 교육정도	3.71
학생 성적 구분	1.27
모 교육정도	1.06
방과 후 학교 참여여부	0.88
부모의 경제활동	0.38
방과 후 학교 참여 시간	0.36
모 연령	0.07

때, 월 평균 총 사교육비는 서울을 기준으로 광역시는 1,200원 낮으며, 중소도시도 1,400원 낮고, 마지막으로 읍면 지역은 2,200원 낮다($e^0_{\text{서울}} - e^{-0.1333}_{\text{광역시}} = 0.12$, $e^0_{\text{서울}} - e^{-0.1541}_{\text{중소도시}} = 0.14$, $e^0_{\text{서울}} - e^{-0.2457}_{\text{읍면지역}} = 0.22$). 또한 학교 급에 따른 총 사교육비의 차이를 보면, 초등학교의 월 평균 총 사교육비를 기준으로 중학교는 2,600원 높으며, 일반계 고등학교는 6,800 원, 전문계 고등학교는 1,500 원 더 높아 일반 고등학생 자녀를 둔 가정이 가장 많은 총 사교육비를 지출하는 것을 알 수 있다($e^0_{\text{초등}} - e^{0.2282}_{\text{중등}} = -0.26$,$e^0_{\text{초등}} - e^{0.5160}_{\text{일반고등}} = -0.68$, $e^0_{\text{초등}} - e^{0.1417}_{\text{전문고등}} = -0.15$). 일반교과 사교육시간에 대해서는 월 단위로 1시간이 증가할수록 총 사교육비는 200원 증가하며, 예체능교과 사교육시간에 대해서는 100원 이하의 증가로 거의 증가하지 않는 것을 알 수 있다($e^{0.0167}_{\text{일반교과 시간}} = 1.02$, $e^{0.0088}_{\text{예체능 교과 시간}} = 1.01$). 학생 성적이 상위 1% 10% 에서부터 10% 단위로 감소함에 따라 월 평균 총 사교육비는 300원씩 감소한다. 그리고 아버지의 교육정도에 대해서는 중졸을 기준으로 증가 할 때 마다 500 원씩 증가되며, 어머니의 연령이 높아짐에 따라서 300 원씩 증가한다. 또한, 어머니의 교육정도도 마찬가지로 학력이 높아짐에 따라 600원씩 증가한다($e^{-0.0275}_{\text{성적}} = 0.97$, $e^{0.0447}_{\text{부 교육정도}} = 1.05$, $e^{0.0273}_{\text{모 연령}} = 1.03$, $e^{0.0583}_{\text{모 교육정도}} = 1.06$). 부모님의 경제활동에 대해서 보면, 아버지 외벌이를 기준으로 어머니 외벌이일 경우에는 월 평균 총 사교육비는 700원 감소하며, 두 분 모두 맞벌이일 경우에는 500원 감소하고, 마지막으로 두 분 모두 경제활동에 참여 하지 않으시면 아버지 외벌이에 비해 약 900원 감소한다($e^0_{\text{아버지외벌이}} - e^{-0.0696}_{\text{어머니외벌이}} = 0.07$, $e^0_{\text{아버지외벌이}} - e^{-0.0466}_{\text{맞벌이}} = 0.05$, $e^0_{\text{아버지외벌이}} - e^{-0.0928}_{\text{모두 참여안함}} = 0.095$). 가구 소득에 대해서는 월 평균 가구 소득이 100만원 미만에서부터 100만원 단위로 증가함에 따라서 800원 증가하며, 방과 후 학습에 참여하는 학생이 참여 하지 않는 학생에 비해 600원이 감소한다. 또 그 참여 시간에 따라서 월단위당 1시간이 증가할 때 마다 100원 미만의 감소로 사실상 변화는 없다 ($e^{0.0734}_{\text{가구소득}} = 1.08$, $e^{-0.0023}_{\text{방과후 학교 참여시간}} = 0.99$, $e^0_{\text{방과후 학교 미참여}} - e^{-0.0574} = 0.06$).

3가지 방법에 대한 비교하기 위해서 linear regression 방법으로 중요변수를 얻어 보았다. Table 4.5을 보면 월 평균 총 사교육비를 설명하는데 있어서 일반교과 사교육시간이 50% 가까이 되는 설명을 하고 있어 총 사교육비를 설명하는데 일반교과 사교육시간 변수에 의존하는 경향이 보인다. 선택된 중요 변수는 일반교과 사교육시간, 학교 급 구분, 지역 구분, 예체능교과 사교육시간, 가구 소득 순서이다.

Random forests 방법으로 총 사교육비를 예측하는데 있어서 일반교과 사교육시간이 36%이상의 중요도를 차지하고 있다 (Table 4.6). 그 뒤로 예체능교과 사교육시간과 가구 소득, 학교 급 구분 등이 중요변수로 선택되었다. 따라서 random forests 방법 역수 총 사교육비를 설명하는데 일반교과 사교육시간에 많이 의존하고 있다. Gradient boosting 방법을 이용해 월 평균 총 사교육비를 설명하는데 있어서 일반

Table 4.6. Important variables in a Random forests model

변수	변수 중요도(%)
일반교과 사교육시간	63.29
학교 급 구분	10.69
예체능교과 사교육시간	10.45
가구 소득	9.34
지역 구분	2.82
방과 후 학교 참여시간	0.97
모 교육정도	0.9
부 교육정도	0.78
부모의 경제활동	0.25
방과 후 학교 참여여부	0.24
성적 구분	0.19
모 연령	0.07

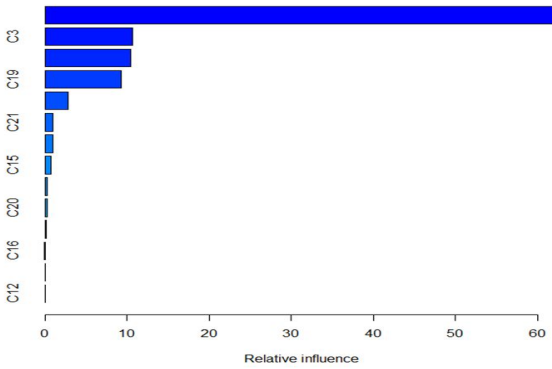

Table 4.7. Important variables in three models

순위	linear regression model	Random Forests	Gradient Boosting
1	일반교과 사교육시간	일반교과 사교육시간	일반교과 사교육시간
2	학교 급 구분	예체능교과 사교육시간	학교 급 구분
3	지역 구분	가구 소득	예체능교과 사교육시간
4	예체능교과 사교육시간	학교 급 구분	가구 소득
5	가구 소득	방과 후 학교 참여시간	지역 구분

교과 사교육시간이 가장 중요한 변수로 선택되었고, 그 뒤로 학교 급 구분과 예체능교과 사교육시간이 선택 되었다. 또한 학생 성적과 어머니와 아버지의 연령과 학생성별에 대해서는 변수 중요도가 0 으로 월 평균 총 사교육비를 설명하는데 중요하지 않음을 알 수 있다.

Linear regression, random forests, gradient boosting 3가지 방법으로 월 평균 총 사교육비를 설명하는데 중요변수에 대해서 비교해 보도록 한다. Table 4.8을 보면 알 수 있듯이 각 모형에 따라서 얻어진 중요 변수가 다르다. 세 가지 방법의 공통점은 월 평균 총 사교육비에 영향을 미치는 가장 중요한 변수가 일반교과 사교육시간이라는 점이다. 또한 변수의 중요도의 차이는 있지만, 중요한 5가지의 변수를 선택하는데 있어서는 일반교과 사교육시간, 예체능교과 사교육시간, 학교 급 구분, 가구 소득의 4개를 공통적으로 선택하였다. 차이점은 linear regression model, gradient boosting 방법에 대해서는 지역 구분이 중요변수로 선택되었지만. random forests 방법에 대해서는 방과 후 학교 참여 시간이 선택되었다.

일반교과 사교육비의 경우에도 3가지 모형은 총 사교육비와 비슷한 결과를 가져왔다. 모형에서 가장 중요한 변수는 일반교과 사교육시간이었고, 그 다음에 학교 급 구분, 지역 구분, 가구 소득 등의 변수가 차례로 선택되었다. 예체능 사교육비의 경우에는 예체능 사교육 시간, 학교 급 구분, 가구 소득, 일반교과 사교육 시간이 중요변수로 선택되었다. 이와 같이 일반교과와 예체능 사교육비의 경우에 선택된 변수들은 비슷하지만 모형이 아주 비슷한 것은 아니다. 실제 일반교과 사교육비의 linear regression model의 계수와 예체능교과 사교육비의 linear regression model의 계수 결과를 토대로 일반교과와 예체능교과 사교육비에 영향을 미치는 변수를 비교해 보면 다음과 같은 사실을 알 수 있다. 먼저, 학교 급 구분 변수 가운데 중학교의 경우에 일반교과는 양수 계수지만 예체능교과는 음수 계수이다. 따라서 초등학교보다 중학교에 진학하면서 일반교과에 대한 사교육은 증가하지만 예체능교과에 대해서는 감소함을 알 수 있다. 또한, 일반교과 사교육비를 설명하는데 있어서 예체능교과 사교육시간이 음수 계수이며, 예체능교과 사교육비를 설명할 때는 일반교과 사교육시간의 계수가 음수이다. 따라서 한쪽 사교육비를 증가시

Table 4.8. Important variables when the total private education time is excluded in the analysis

순위	linear regression model	Random Forests	Gradient Boosting
1	학교 급 구분	가구 소득	학교 급 구분
2	가구 소득	방과 후 학교 참여시간	가구 소득
3	지역 구분	학교 급 구분	지역 구분
4	부의 교육정도	학생 성적 구분	방과 후 학교 참여시간
5	방과 후 학교 참여여부	지역 구분	학생 성적 구분

키기 위해서는 다른 쪽 사교육비는 감소시켜야 한다는 것을 유추할 수 있다. 성별에 대해서도 차이가 나는데 이때 일반교과에서는 남학생의 경우 사교육비가 증가하며, 예체능교과에서는 여학생의 경우 사교육비가 높다. 또한, 성적에 대해서도 반대의 결과가 나타남을 알 수 있다. 일반교과의 경우 성적이 높을수록 일반교과 사교육비가 높은데 반해 예체능교과는 성적이 낮을수록 예체능교과 사교육비가 증가한다. 마지막으로 일반교과 사교육비에서는 부모님의 경제활동이 최종 모형에 선택되었고, 계수도 유의한 반면, 예체능교과 사교육비를 설명하는데 있어서는 최종 모형에 선택되지 않았다.

B. 사교육시간을 분석에 포함하지 않은 경우

앞에서 볼 수 있듯이 사교육시간이 모든 모형에서 가장 중요한 변수로 선택되었다. 이는 어찌 보면 당연한 일이지만 우리는 사교육 시간을 배제한 경우에 모형의 설명력과 다른 변수들의 기여도가 어떻게 변하는 지가 궁금하였다. 사교육 시간을 모형에 포함하지 않고 분석한 결과 중요변수로 선택된 것은 다음과 같다 (총 사교육비).

Table 4.8에서 볼 수 있듯이 공통으로 선택된 변수들은 학교 급 구분, 가구 소득, 방과 후 학교 참여시간, 지역구분 이다. linear regression model에서는 부의 교육 정도가 중요변수 중에 하나였고, random forests와 gradient boosting에서는 학생 성적 구분이 중요변수로 선정되었다. 일반교과 사교육비의 경우에는 순위가 조금 다른 것을 제외하고는 거의 동일한 변수들이 중요변수로 선택되었고 예체능 사교육비의 경우도 거의 유사하나 학생 성별과 모의 교육정도가 중요변수로 선택된 차이가 있었다. 우리는 일반교과 와 예체능교과 사교육비의 linear regression model에서 회귀계수를 이용하여 두 그룹을 비교하였다. 그 결과, 초등학교를 기준으로 중학교가 되면 일반교과 사교육비는 양수 계수로 증가 하지만 예체능교과 사교육비에 대해서는 음수 계수로 감소한다. 따라서 초등학교까지는 예체능 사교육을 시키지만 중학교로 들어가면서 예체능 사교육을 줄이고 일반교과 사교육을 늘리는 현상을 알 수 있다. 그리고 학생에 관한 변수에서도 차이가 나는데, 학생의 성별에 따라서 일반교과 사교육의 경우 남학생이 더 높고, 예체능교과 사교육의 경우에는 여학생의 사교육비가 더 높았다. 또한, 성적에 따라서도 다른 결과를 얻을 수 있었다. 일반교과의 경우 성적이 상위권일수록 사교육비가 높았으며, 예체능교과의 경우 성적이 하위권일수록 사교육비가 높았다.

5. 결론

본 논문에서는 2012년 초, 중, 고등학교 학생의 사교육비에 영향을 미치는 중요 변수와 이 변수들이 어떻게 사교육비에 영향을 미치는 지를 알아보았다. 우선 사교육비 응답에 있어서 0원에 대한 비율을 대표적인 변수 별로 알아보았다. 사교육비 0원에 대한 응답을 지역에 따라서 보면, 읍면지역, 중소도시, 광역시, 서울 순서로 높았다. 또한 학교 급에 따르면 전문계 고등학교가 가장 많았고, 그 다음으로 일반계 고등학교, 중학교, 초등학교 순서였으며, 성적에 따라서는 상위에서 하위로 갈수록 사교육비 0원에 대한 비율이 늘어났다. 또한 아버지의 연령이 높을수록 높았고, 아버지의 교육 수준이 낮을수록 사교육비 0 원에 대한 비율이 높았다. 그리고 사교육비에 영향을 미치는 요인을 밝히고자 실시한 분석 중

첫 번째로 분류 분석을 실시하였다. 그 결과 일반 및 예체능교과 사교육시간, 학교 급 구분, 가구 소득, 학생 성적 등이 중요변수로 선택되었다. 이때 일반 및 예체능교과 사교육시간은 분류분석에서 중요변수인 것은 너무나도 당연한 결과이고, 학교 급 구분 변수는 전문 고등학교의 경우 사교육을 하지 않는 응답이 많고, 학생 성적이 낮을수록, 가구 소득이 낮을수록 사교육비 0원의 비율이 높아서 선택된 결과라고 판단한다. 그다음 분석으로 사교육비 0원이 아닌 자료, 즉 사교육을 시행하고 있는 학생을 대상으로 세 가지의 다른 회귀 분석의 방법을 적용하여 사교육비에 영향을 미치는 중요변수에 대해 알아보았다. 총 사교육비를 설명하는데 중요 요인은 일반교과 사교육시간, 학교 급 구분, 지역 구분, 예체능교과 사교육시간, 가구 소득 순 이였다. 총 사교육비는 일반교과 사교육비와 양상이 비슷하였다. 이는 총 사교육비를 구성하는데 있어서 일반교과 사교육비가 많은 부분을 차지하고 있기 때문이다. 예체능교과 자료분석의 결과로 얻은 중요변수는 예체능교과 사교육시간, 학교 급 구분, 가구 소득, 일반교과 사교육시간이다. 예체능교과 사교육시간과 가구 소득이 증가할수록 예체능교과 사교육비는 증가하고, 일반교과 사교육시간이 증가할수록 예체능교과 사교육비는 감소하였다. 또한, 학교 급 구분에 따른 비교는 일반 고등학교, 전문 고등학교, 초등학교, 중학교 순서이다. 중요변수는 linear regression model 과 random forests, 그리고 gradient boosting 에서 약간 차이가 났다. Decision tree는 사실 아주 비선형 모델이고 이런 tree를 사용하는 random forests와 gradient boosting 방법론도 비선형일 수밖에 없으므로 linear regression model과 차이가 날 수 밖에 없다고 생각한다. 이번 연구를 통하여 얻은 결론으로 볼 때, 우리나라 초, 중, 고 학생의 사교육은 일반교과 사교육시간, 예체능교과 사교육시간, 학교 급 구분, 지역 구분, 가구 소득 등에 많은 영향을 받고 있다. 여기서 학교 급 구분 변수에 대해서는 '초등학교에는 예체능교과에 대한 사교육을 많이 시행한다.'와 같은 학교의 급이 오를수록 변하는 사교육비 경향을 나타낼 뿐 외부적인 사교육비 증가, 감소의 영향이 되진 않았다. 하지만 지역 구분, 가구 소득에 따라서 사교육비가 다르다는 점은 모든 학생들이 동등한 교육의 기회를 제공받지 못한다는 점에서 우리 사회가 함께 해결해야 할 문제점이라고 생각한다.

References

김진희 (2010). 2008년 한국 초중고 학생에 대한 사교육비 자료분석, *이화여대 석사논문*.

박창이 외 (2011). *R을 이용한 데이터 마이닝*. 교우사.

김수현 (2010). 사교육비 결정요인 분석으로 본 사교육 실태, *새로운 사회를 여는 연구원*.

윤영회, 장수안 (2013). 2012년 사교육비조사 결과, *통계청 사회통계국 복지통계과*.

Breiman, L. (2001). Random forests. Machine Learning, **45**, 5-32.

Friedman, J. H. (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367-378, http://statweb.stanford.edu/ jhf/ftp/stobst.pdf

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, NY, USA.

Ridgeway, G. (2012). Generalized Boosted Models: A guide to the gbm package.

Venables, W.N. and Ripley, B. D. (2003). *Modern Applied Statistics with S. Springer*, NY, USA.

초,중,고 사교육비 영향요인 분석

이혜정[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2014년 09월 16일 접수, 2014년 10월 14일 수정, 2014년 11월 06일 채택)

요 약

요약 본 논문에서는 현재 우리 사회에서 큰 이슈가 되고 있는 사교육비에 영향을 미치는 요인을 분석하였다. 사교육비가 논란이 되는 이유 가운데 하나는 각 가정에서 소비되는 전체 교육비 가운데 사교육비가 차지하는 비중이 높으면 경제적으로 교육에 대한 기회균등이 이루어지지 않는 문제를 유발할 수 있기 때문이다. 또한, 과도한 사교육으로 인해 기본적인 학교 교육보다 더 많은 시간과 비용을 할애한다면, 계층 간, 지역 간의 양극화와 더불어 학교의 기능이 저하되는 요인으로 작용할 수 있기 때문이다. 우리는 다양한 회귀분석 방법론과 분류분석 방법론을 이용하여 사교육비에 영향을 미치는 중요변수가 무엇인지를 파악하고 또 각각의 변수들이 어떻게 사교육비에 영향을 미치는 지를 분석하였다. 그 결과 대도시가 중소도시 보다 사교육비 지출이 높았고, 일반 고등학교, 중학교, 전문 고등학교, 초등학교 순으로 사교육비 지출이 높았다. 그리고 가구소득이 높을수록 사교육비 지출도 높았다.

주요용어: 주요용어: 사교육비, 회귀분석, 분류분석, random forests, gradient boosting.

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A2012817).

[1]교신저자: (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 부교수.
E-mail: josong@ewha.ac.kr