서울시립대학교
UNIVERSITY OF SEOUL

학기말 보고서

로지스틱 회귀분석을 통한
대학원 합격여부 예측

과목명	회귀분석 II
교수명	김규성 교수님
학 과	통계학과
학 번	
이 름	이유민
제출일	2020. 12. 09

차례

---

1. 서론

1.1 연구 배경 및 연구 목적

대학원 입학자 수는 최근 3년 새 늘어나고 있다. 대학정보공시 자료에 따르면 전국 일반·특수대학원 입학자 수는 2017년 10만7371명에서 올해 11만1843명으로 4472명(4.1%) 증가했다. 같은 기간 일반대학원 입학자 수는 6만2791명에서 6만6876명으로 4085명(6.5%) 늘었다. 특히 코로나19 사태가 장기화되면서 대학원에 진학하길 희망하는 대학생들도 점점 늘어나는 추세이다.

이러한 상황에서, 대학교 입시와 대학원 입시는 결이 많이 다르다. 담임 선생님, 부모님과의 상담을 통해 대학교를 결정했던 것과는 상반되게, 대학원 진학은 풍부한 양질의 정보가 제공되지 않기 때문이다.

이러한 점을 노려 입시 컨설팅 회사들은 대학원 입학 준비생들을 신규 `고객`으로 받아들이려는 움직임을 보인다. 이제는 고3 수험생들을 대상으로 한 컨설팅 서비스가 국내 대학원 입학을 준비하는 대학생들 사이에서도 서서히 확산하고 있다는 것이다.

하지만 문제는 여기에서 발생한다. 단순히 입학에 대한 정보를 제공하는 것이 아니라, 업체들이 대학원 입학 준비생들에게 불안감을 부추기며 회사의 서비스를 이용하게 유도한다는 것이다. 자기소개서 작성부터 면접 대비까지 책임지는, 즉 `패키지`를 구입하는 경우 대학원 한 곳당 준비 비용은 약 300만원 정도이다. 좀 더 구체적으로 말하자면, 서울대 대학원은 385만원, 고려대·연세대는 각각 330만원, 성균관대·서강대 등은 각각 275만원 등이다.

이번 분석에서는 대학원 합격과 관련한 데이터를 통해 변수들 사이에 어떤 관계가 존재하는지 살펴본다. 또한 로지스틱 회귀분석을 통한 대학원 합격 여부를 예측하며 유의한 요소들을 찾으려고 한다. 이 결과는 앞으로 대학원에 진학하길 희망하는 대학생들에게 도움이 될 수 있는 연구의 출발점이 될 수 있을 것이다.

1.2 문헌 연구

1.2.1 4년제 대졸자의 대학원진학 결정요인 분석

본 연구에서는 4년제 대학 졸업생의 진로결정 및 대학원 진학 선택에 영향을 미치는 요인 분석하고 있다. 결론은 크게 세 부분으로 나눌 수 있는데, 대학원 진학자는 전반적으로 나이가 어리고 부모의 교육수준이 높으며, 이공계 등 특정전공자의 대학원 진학률이 높았다. 대학원 진학자는 학부 당시 전공커리큘럼이나 교수진의 전공능력과 열의가 중요한 영향을 미치고 있었으며, 대학 평점이나 영어점수를 주요 진학 준비 행동으로 실시하고 있었다. 마지막으로 대학원 유형에 따른 차이를 살펴본 결과, 전문대학원은 부모학력이 높은 서울권 대졸자 중 사회과학 전공자가 진학하는 비중이 높고, 일반대학원은 이공계 전공자, 특수대학원은 인문계열과 예체능계 전공자가 상대적으 로 진학하는 특성을 보였다. 해당 논문에서 우리나라 대학원 선택이 어떠한 특징을 가지는지를 참고하여 본 분석에서도 보다 유의미한 결과를 이끌어 낼 수 있을 것이다.

1.2.2 USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

해당 논문에서는 데이터 마이닝을 사용하여 학생들의 성과를 예측하고 있다. 이번 연구에서 다루는 데이터의 raw data인 UCL의 데이터를 참고하여 진행한 연구로, 회귀분석을 통해 얻은 결과들을 해석하는 데에 있어 도움을 줄 것이다.

1.2.3. 대학원의 교육/연구/ 경쟁력 확보 방안

해당 논문에서는 우리나라 대학원의 교육과 연구 경쟁력, 대학원 과정 운영 전반의 효율성과 책무성을 제고하기 위한 연구를 목적으로 한다. 현재 대학원 과정 제도 전반에 대해 진단한 후, 대학원 교육과 연구 경쟁력을 위한 제도 개선방안을 제시한다. 이 연구를 통해 전반적인 대학원 과정이 어떻게 진행되는지에 대한 정보를 얻고, 대학생들이 대학원 선택에 있어서 중요하게 생각하는 요소가 무엇인지에 대해 참고하여 이번 분석에 녹여 내려 한다.

1.3 데이터 설명

1.3.1 데이터명

: Graduate Admission 2

1.3.2 출처

: kaggle (주소: https://www.kaggle.com/mohansacharya/graduate-admissions)

: UCL Machine Learning Repository (raw data)

(주소: https://archive.ics.uci.edu/ml/datasets/student+performance)

1.3.3 description

본 데이터는 인도에서의 대학원 입학 합격 예측을 위해 구축된 데이터셋이다. 본 데이터는 UCLA 대학원 데이터를 기반으로 하고 있으며, 데이터에는 대학원에 진학 시에 영향을 줄 법한 설명변수 6개와 함께 대학원에 진학할 확률을 뜻하는 변수를 담고 있다. kaggle에 여러 버전의 graduate admission data가 존재했기 때문에 동일 column을 기준으로 데이터를 병합한 후 중복 데이터를 삭제하여 최종 dataset을 구축하였다.

Raw data의 컬럼들은 다음과 같다 (총 7열).

1. GRE Scores (340점 만점)

: Graduate Record Examination 변수로, 미국을 비롯한 여러 영어권 국가들의 대학원 및 경영대학원에 입학하려는 학생들을 평가하는 시험 점수 변수이다.

2. TOEFL Scores (120점 만점)

: Test of English as a Foreign Language, 즉 토플 점수 정보를 담은 변수이다.

3. University Rating (5점 만점)

: 대학 평가 점수로, 낮은 대학이면 1, 높은 대학일수록 5점에 가까운 점수를 부여한다.

4. Statement of Purpose and Letter of Recommendation Strength (5 점 만점)

: 목적 진술서(Statement of Purpose)와 추천 정도(Letter of Recommendation Strength)를 의미한다.

5. Undergraduate GPA (10 점 만점)

: 학부 평점 정보를 담고 있는 변수이다.

6. Research Experience (0 또는 1): 연구 경험

: 학부 시절 연구 경험 유무를 나타낸다. 연구 경험이 있는 경우는 1, 그렇지 않으면 0을 가지는 바이너리 변수이다.

7. Chance of Admit (0 ~ 1 범위)

: 대학원에 합격할 확률을 담은 변수로, 합격과 불합격을 예측하기 위한 타겟변수를 생성하는 데에 쓰이는 변수이다.

로지스틱 회귀분석을 수행하기 위해 Chance of admit 변수의 분포를 plot을 통해 확인한 후, 중위수가 존재할 것으로 예상되는 0.7을 기준으로 하여 admit>0.7인 경우 1, 아닌 경우 0의 값을 가지는 result 변수를 생성하였다.

1.4 결과 활용 및 기대 효과

분석 결과는 일차적으로 생각했던 것과 같이, 대학원 진학을 희망하는 대학생들이 본인에게 잘 맞는 동시에 합격 가능성이 높은 대학에 입학할 수 있도록 도울 수 있을 것이다. 또한 대학 졸업자의 취업 뿐만 아니라 대학원 진학을 고려함으로써 대학생의 진학과 관련한 결정을 미리 예측하고, 대학생들에게 좀 더 체계적이고 적절한 진로 서비스를 제공하는데에 있어서 중요한 정보가 될 것이다. 또한, 최근 교육을 인적자원에 대한 투자개념으로 받아들이는 사회의 시각에 따라, 대한민국의 고급인력으로 성장할 가능성을 가진 인적자원을 사전에 파악하고 이에 대한 지원을 강화하는 방안들을 개발하는데 도움을 줄 수 있을 것이다.

2. 본론

2.1 분석 방법 소개

대학원 합격에 영향을 미치는 요소를 분석하기 위해 반응변수를 이산형인 Result (대학원 합격 여부)로 설정했다. 이와 같은 구조에서 대표적인 회귀분석 방법으로는 로지스틱 회귀분석이 있다.

로지스틱 회귀분석은 설명변수가 연속형 반응변수에 영향을 미치는 정도를 살펴보는 것에서 일반 선형 회귀분석과 비슷하다. 하지만 일반 선형회귀분석에 사용되는 반응변수는 연속형 변수이며, 정규분포를 따른다는 가정이 필요하지만, 로지스틱 회귀분석은 반응변수가 이산형이므로 확률모형을 가질 수 없기 때문에 선형회귀분석 보다는 만족해야 하는 가정을 필요로 하지 않는 장점을 가지고 있다. 앞서 언급했듯이 로지스틱 회귀는 일반 선형모델의 특수한 경우로 볼 수 있으므로 선형 회귀와 유사하다.

하지만 로지스틱 회귀의 모델은 이산형 모델을 사용하였을 때 반응 변수 y의 결과가 범위 [0,1]로 제한되고 반응변수가 이진 형태이기 때문에 조건부 확률의 분포가 정규분포 대신 이항분포를 따른다는 점에서 차이를 갖고 있다. 그러므로 반응변수의 결과가 0과 1, 두 개의 경우만 존재하는데, 단순 선형 회귀를 적용하면 범위를 벗어나는 결과가 나올 수 있기 때문에 적절하지 않다. 이를 해결하기 위해 로지스틱 회귀는 연속이고 증가함수이며 [0,1]에서 값을 갖는 다음과 같은 모형을 사용한다.

$$logit = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$
$$where\ logit = \log\left(\frac{p}{1-p}\right), p = \Pr(Y = 1|X)$$

이번 분석에서는 admission 데이터의 result를 반응변수로 설정하였다. 또한 나머지 8개의 변수 중 데이터의 index 정보를 담은 no를 제외한 7개의 변수를 설명변수로 설정하여 어떤 변수가 대학원 합격 여부에 더 큰 영향을 미치고 있는지를 분석해보고자 한다.

로지스틱 분석을 기반으로, 입학 예측 변수를 타겟변수로 하였다. 입학 예측 변수(result)는 데이터 전처리 과정에서 설명한 것처럼, (0,1) 사이의 값을 가지는 대학

원 입학 확률 변수 admit의 값이 0.7 이상인 경우 1(대학원에 진학함), 미만인 경우 0(대학원에 진학하지 않음)으로 생성하였다. 0.7을 기준으로 한 것은 산점도로 확인한 admit 변수의 중앙값이 0.7 언저리에 존재했기 때문이다.
분석을 위해 사용한 프로그램은 파이썬과 SAS 9.4이다. 기본적인 데이터 EDA와 변수 생성은 파이썬에서 진행하였으며, 본격적인 분석은 모두 SAS로 진행하였다.
Wald 검정 통계량과 변수 선택 방법 중 Stepwise Selection을 통해 유의한 설명변수가 무엇인지 확인해보았다. stepwiseselection은 모든 변수가 포함된 모델에서 출발하여 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져 있는 변수 중에서 기준 통계치를 가장 개선 시키는 변수를 추가한다. 그리고 이러한 변수의 추가 또는 삭제를 반복한다. 또한 우도비통계량, Hosmer and Lemeshow Goodness-of-Fit Test과 deviance 등을 사용하여 모형 적합을 확인해보았다.

데이터 전처리에 있어 필수적인 과정인 결측값 처리, 이상치 처리, 데이터셋을 병합하는 과정에서 조절해야 했던 여러 과정들은 이번 분석에서 중요하게 다룰 필요가 없으므로 짧게 줄인다.

먼저, 최종적으로 생성한 데이터는 다음과 같다. 첫 10개의 행만 나타내었다.

OBS	no	gre	toefl	univ	sop	lor	cgpa	research	admit	result
1	1	337	118	4	4.5	4.5	9.65	1	0.92	1
2	2	324	107	4	4	4.5	8.87	1	0.76	1
3	3	316	104	3	3	3.5	8	1	0.72	1
4	4	322	110	3	3.5	2.5	8.67	1	0.8	1
5	5	314	103	2	2	3	8.21	0	0.65	0
6	6	330	115	5	4.5	3	9.34	1	0.9	1
7	7	321	109	3	3	4	8.2	1	0.75	1
8	8	308	101	2	3	4	7.9	0	0.68	0
9	9	302	102	1	2	1.5	8	0	0.5	0
10	10	323	108	3	3.5	3	8.6	0	0.45	0

2.2 데이터 분석 및 결과 설명

2.2.1 변수 빈도 및 기초통계량 확인, 상관계수 확인

데이터는 총 500 행이며, 반응변수에 해당하는 result변수가 1(대학원 합격)인 사람은 287명, result 변수가 0 (대학원 불합격)인 사람은 213명이다.

result	빈도	백분율	누적 빈도	누적 백분율
0	213	42.60	213	42.60
1	287	57.40	500	100.00

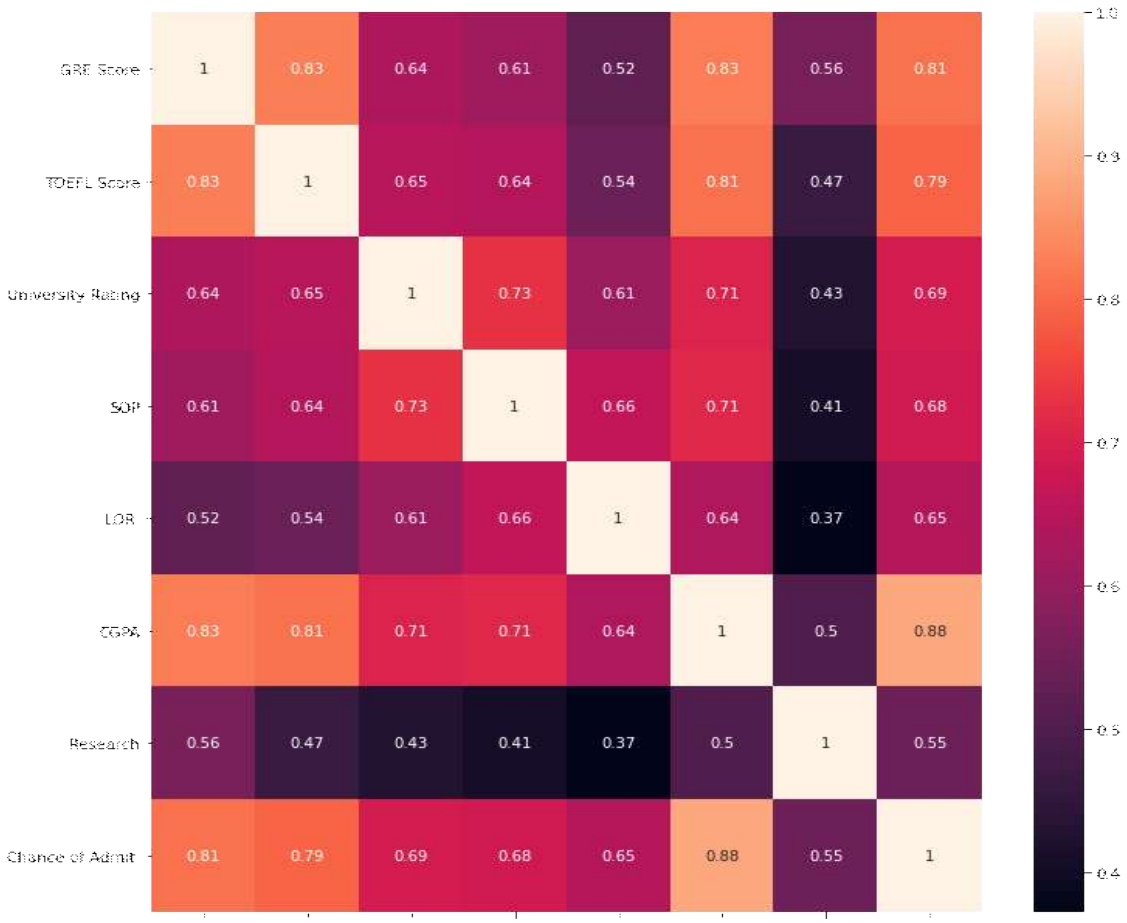
설명변수들이 어느 정도 수준의 값을 가지는지를 보기 위해 proc means 프로시저로 확인하였다.

변수	N	평균	표준편차	최솟값	최댓값
no	500	250.50	144.48	1.00	500.00
gre	500	316.47	11.30	290.00	340.00
toefl	500	107.19	6.08	92.00	120.00
univ	500	3.11	1.14	1.00	5.00
sop	500	3.37	0.99	1.00	5.00
lor	500	3.48	0.93	1.00	5.00
cgpa	500	8.58	0.60	6.80	9.92
research	500	0.56	0.50	0.00	1.00
admit	500	0.72	0.14	0.34	0.97
result	500	0.57	0.49	0.00	1.00

또한 변수들 사이의 상관성이 존재하는지를 알아보기 위해 상관관계도 살펴보았다. SAS와 파이썬에서도 산점도를 구하였으며, 이산형 변수를 포함하기 때문에 파이썬에서는 spearson rank correlation coefficient 옵션을 지정해주었다.

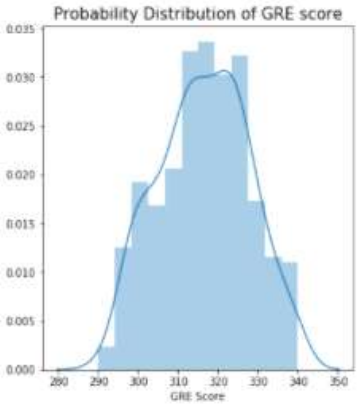
피어슨 상관 계수, N = 500
H0: Rho=0 가정하에서 Prob > |r|

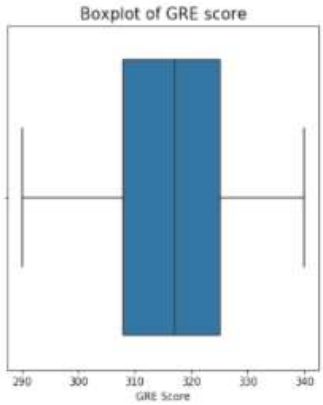
	no	gre	toefl	univ	sop	lor	cgpa	research	admit	result
no	1.00000	-0.10384 0.0202	-0.14170 0.0015	-0.06764 0.1309	-0.13735 0.0021	-0.00369 0.9343	-0.07429 0.0971	-0.00533 0.9053	0.00851 0.8495	0.01784 0.6907
gre	-0.10384 0.0202	1.00000	0.82720 <.0001	0.63538 <.0001	0.61350 <.0001	0.52468 <.0001	0.82588 <.0001	0.56340 <.0001	0.81035 <.0001	0.67943 <.0001
toefl	-0.14170 0.0015	0.82720 <.0001	1.00000	0.64980 <.0001	0.64441 <.0001	0.54156 <.0001	0.81057 <.0001	0.46701 <.0001	0.79223 <.0001	0.63166 <.0001
univ	-0.06764 0.1309	0.63538 <.0001	0.64980 <.0001	1.00000	0.72802 <.0001	0.60865 <.0001	0.70525 <.0001	0.42705 <.0001	0.69013 <.0001	0.57102 <.0001
sop	-0.13735 0.0021	0.61350 <.0001	0.64441 <.0001	0.72802 <.0001	1.00000	0.66371 <.0001	0.71215 <.0001	0.40812 <.0001	0.68414 <.0001	0.56648 <.0001
lor	-0.00369 0.9343	0.52468 <.0001	0.54156 <.0001	0.60865 <.0001	0.66371 <.0001	1.00000	0.63747 <.0001	0.37253 <.0001	0.64536 <.0001	0.53849 <.0001
cgpa	-0.07429 0.0971	0.82588 <.0001	0.81057 <.0001	0.70525 <.0001	0.71215 <.0001	0.63747 <.0001	1.00000	0.50131 <.0001	0.88241 <.0001	0.70147 <.0001
research	-0.00533 0.9053	0.56340 <.0001	0.46701 <.0001	0.42705 <.0001	0.40812 <.0001	0.37253 <.0001	0.50131 <.0001	1.00000	0.54587 <.0001	0.49116 <.0001
admit	0.00851 0.8495	0.81035 <.0001	0.79223 <.0001	0.69013 <.0001	0.68414 <.0001	0.64536 <.0001	0.88241 <.0001	0.54587 <.0001	1.00000	0.80836 <.0001
result	0.01784 0.6907	0.67943 <.0001	0.63166 <.0001	0.57102 <.0001	0.56648 <.0001	0.53849 <.0001	0.70147 <.0001	0.49116 <.0001	0.80836 <.0001	1.00000

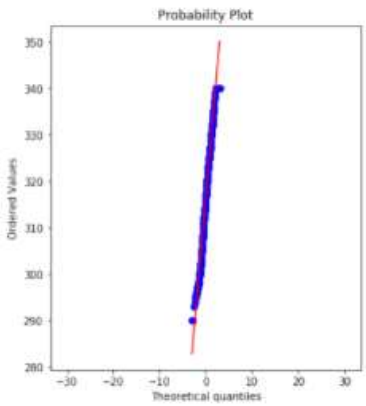


result에 대해 gre, cgpa등의 변수가 상관이 있는 것처럼 보인다.

또한 GRE Score를 포함하여 대부분의 변수들이 정규성을 따르고 있어 별도의 변환을 할 필요는 없다고 생각했다. 예시로, 하단의 세 플롯은 GRE Score의 probability distribution, violin plot, Q-Q plot을 나타내었다. Outlier가 존재하지 않고 정규성을 보이는 것을 확인할 수 있었다.







외에도 이상치로 보이는 행 4개 정도를 발견했으나, 하나씩 살펴 본 결과 정상적인 값을 가지는 행이였기 때문에 이번 분석은 데이터의 500개 행을 그대로 사용한다.

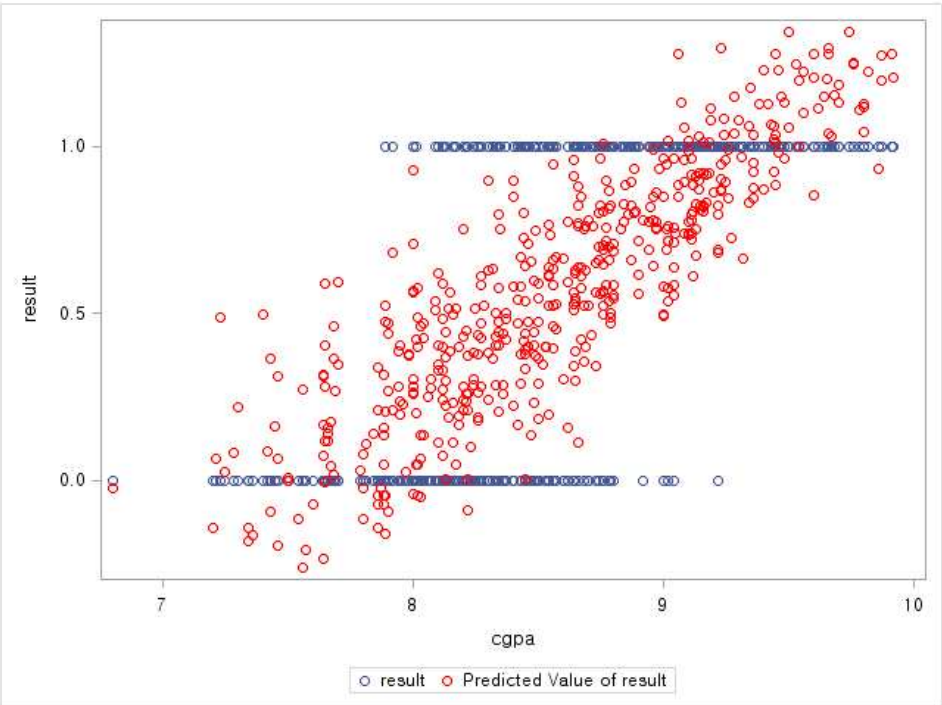
2.2.2 선형 회귀모형 적합

먼저, 선형회귀분석을 적합하였다. result를 반응변수, 나머지 반응변수 중 gre와 lor을 각각 설명변수로 놓고 반응변수와 선형 회귀분석을 진행한 후, 예측값과 실제 값을 함께 나타낸 결과는 다음과 같다.

유의수준 0.05 하에서 gre와 lor은 모두 유의함을 확인할 수 있다. 하지만 회귀선에서 Result의 반응변수는 0과 1의 값만을 가지는데 회귀직선들이 이 범위를 벗어난 값들을 가지고 있다. 그러므로 선형회귀분석은 적절한 분석이 아니라고 여겨진다. 따라서, 이러한 데이터에서 대표적으로 사용하는 로지스틱 회귀모델을 적합하였다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.48933	0.48061	-15.58	<.0001
gre	1	0.02400	0.00162	14.81	<.0001
lor	1	0.13433	0.01978	6.79	<.0001



2.2.3 로지스틱 회귀분석 - 1

반응변수 result와 모든 설명변수에 대해 로지스틱 회귀분석을 적합시킨 결과이다.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	391.0040	7	<.0001
Score	275.5346	7	<.0001
Wald	116.7634	7	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.6782	8	0.7914

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	291.1509	492	0.5918	1.0000
Pearson	354.0046	492	0.7195	1.0000

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	57.4723	7.1539	64.5410	<.0001
gre	1	-0.0877	0.0266	10.8754	0.0010
toefl	1	-0.0623	0.0463	1.8122	0.1782
univ	1	-0.4096	0.2052	3.9836	0.0459
sop	1	-0.0473	0.2354	0.0404	0.8406
lor	1	-0.6292	0.2238	7.9040	0.0049
cgpa	1	-2.3251	0.5428	18.3507	<.0001
research	1	-0.8040	0.3134	6.5827	0.0103

먼저, 귀무가설 $H_0 : \beta = 0$ 으로 놓고 우도비 검정통계량(Likelihood Ratio)를 살펴보았다. 검정통계량 값은 391.0040으로 p-value가 0.0001보다 작기 때문에 유의수준 alpha=0.01 하에서 검정 결과가 설명변수들로 반응변수를 잘 설명한다고 볼 수 있다.

다음으로 적합도 검정을 통해 모델이 적합한지 확인해보았다. 이 검정에서는 귀무가설은 '모형이 적합하다' 이므로 유의확률이 높아 귀무가설을 기각하지 못하면 모형이 적합한 것으로 본다. 그러므로 Hosmer and Lemeshow Goodness-of-Fit Test에서는 우리가 적합 시킨 모델이 적합하다는 귀무가설에 대하여 p-value 값이 1로 매우 크므로 모델이 적합하다는 결론을 내릴 수 있다.
Deviance 적합도 검정에서도 p-value 값이 1로 모델이 적합하다는 결론을 내릴 수 있다. Pearson 적합도 검정의 경우에도 유의하다는 결론이 나왔으나, 데이터가 이항 반응변수인 경우 Pearson 적합도 검정은 부정확한 결과값을 가져올 수 있다고 알려져 있기 때문에, 이 검정 결과는 크게 신경쓰지 않기로 한다.
각각의 변수들이 유의한지 보기 위해 Wald 통계량을 사용했다. 이 때, 귀무가설은 회귀계수 $\beta_i = 0 \, (i = 1,2,...,8)$ 이다. 대부분의 변수들은 귀무가설을 기각하여 유의하다고 볼 수 있는데, sop 변수에 대한 Wald 검정 통계량의 p-value 값이 0.8406으로 매우 크므로 유의하지 않다. 그러므로 이 변수를 제거해야 한다고 판단했다.

2.2.4 로지스틱 회귀분석 – 2

이어서, 각 변수들에 대해 살펴본 결과, 유의수준 0.02 하에서 유의하지 않는 몇몇 변수가 보여 해당 변수들을 제거해주려 한다. 이를 뒷받침할 근거를 더 찾기 위해, 변수선택방법 중 stepwise selection을 이용하였다. Forward와 backward로 진행한 결과도 동일했으며, 실행 결과는 다음과 같다.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	389.0440	5	<.0001
Score	275.2057	5	<.0001
Wald	118.3332	5	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.1428	8	0.7422

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	293.1108	494	0.5933	1.0000
Pearson	353.4841	494	0.7156	1.0000

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	56.7643	6.9767	66.1993	<.0001
gre	1	-0.1013	0.0246	16.9481	<.0001
univ	1	-0.4687	0.1863	6.3300	0.0119
lor	1	-0.6367	0.2105	9.1521	0.0025
cgpa	1	-2.5104	0.5229	23.0452	<.0001
research	1	-0.7895	0.3118	6.4110	0.0113

2.2.5 로지스틱 회귀분석 - 3. 최종 결정된 모델

유의수준을 0.02로 하였을 때, 유의한 변수는 gre, univ, lor, cgpa, research이다. 따라서 해당 5개의 변수를 가지고 다시 로지스틱 회귀분석을 진행했다.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	389.0440	5	<.0001
Score	275.2057	5	<.0001
Wald	118.3332	5	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.1428	8	0.7422

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	293.1108	494	0.5933	1.0000
Pearson	353.4841	494	0.7156	1.0000

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	56.7643	6.9767	66.1993	<.0001
gre	1	-0.1013	0.0246	16.9481	<.0001
univ	1	-0.4687	0.1863	6.3300	0.0119
lor	1	-0.6367	0.2105	9.1521	0.0025
cgpa	1	-2.5104	0.5229	23.0452	<.0001
research	1	-0.7895	0.3118	6.4110	0.0113

모델 설명은 2.2.6절에 이어서 진행한다.

2.2.6 적합된 모델 설명 및 오즈비 구하기

Likelihood Ratio(우도비 검정통계량)은 389.0440이며 P-value의 값이 0.0001보다 작기 때문에 유의수준 0.02 하에서 설명변수들이 포함된 모델이 적합하다고 할 수 있다. Hosmer and Lemeshow Goodness-of-Fit Test(적합도 검정) 결과와 Deviance 적합도 검정 결과도 모형이 적합하다는 귀무가설에 대하여 p-value 값이 모두 1로 적합하다고 할 수 있다. 각각의 변수들에 대한 Wald 검정 통계량의 p-value 값도 0.2보다는 작으므로 이 5개의 설명변수들을 포함한 모델이 적합한 모델이라고 판단했다.

SAS 결과값을 바탕으로 추정한 로지스틱 회귀식을 써보면 다음과 같다.

$$logit = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 56.7642 - 0.1013*gre - 0.4687*univ - 0.6367*lor - 2.5104*cgpa - 0.7895*research$$

또한 Odds Ratio를 통해 설명변수들과 반응변수의 관계에 대해 알아보았다.

$$\frac{\frac{\pi(x_i+1)}{1-\pi(x_i+1)}}{\frac{\pi(x_i)}{1-\pi(x_i)}} = e^{\beta_i}$$ 를 이용한다.

또한, 이 때 e^{β_i}의 값은 Point Estimate에 나타나 있다.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gre	0.904	0.861	0.948
univ	0.626	0.434	0.902
lor	0.529	0.350	0.799
cgpa	0.081	0.029	0.226
research	0.454	0.246	0.837

결과를 토대로 분석 결과를 수치적으로 해석해 보자. 나머지 설명변수들이 고정되어 있다는 조건 하에서 gre 점수가 1점 늘어날 때마다 대학 불합격률은 0.904배 낮아진다. 같은 조건으로 해석하면 univ점수가 1씩 증가할 때마다, lor이 1씩 증가할 때마다, 학부 평점이 1 증가할 때마다, 연구 경험이 있을 때마다 대학 불합격률은 각각 0.626, 0.529, 0.081, 0.454배 낮아지는 것을 확인할 수 있다.
오즈비의 신뢰구간을 보면 모든 설명변수들이 1을 포함하고 있지 않기 때문에 각 설명변수들은 반응변수에 영향을 끼친다고 할 수 있다. 구간이 1을 포함한다면 이는 설명변수가 반응변수에 크게 영향을 끼치지 않을 가능성이 있음을 의미한다.

2.2.7 표준화한 변수로 로지스틱 회귀분석

유의미한 결과를 보이는 데이터들이였으나, 각 변수마다 편차가 차이가 심하게 나는 데이터이다. 이처럼 설명변수 간 스케일의 차이 때문에 가장 큰 영향을 미치는 변수를 결론내리기는 어렵다고 판단하였다. 따라서 설명변수 데이터를 표준화시킨 후, 로지스틱 회귀분석을 실시하였다. 그 결과, 설명변수를 표준화한 데이터로 추정된 회귀식은 앞서 진행했던 표준화시키지 않은 데이터를 분석했을 때 나오는 결론과는 다르다는 것을 알 수 있었다. 결과는 하단에 첨부되어 있다.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
gre	0.318	0.185	0.549
univ	0.585	0.385	0.888
lor	0.555	0.379	0.813
cgpa	0.219	0.118	0.407
research	0.676	0.499	0.915

2.3 분석의 타당성 설명

먼저 제일 처음에 진행한 선형회귀분석의 경우, 계수들이 유의하다고 판단되나 산점도로 확인한 결과 반응변수의 특징을 제대로 설명하지 못하고 있으므로 적합하지 않은 모형으로 판단하였다.

이번 분석에서 설정한 반응변수는 0과 1의 값만 가지는 이항변수이므로 다른 회귀분석 방법보다 로지스틱 회귀분석으로 진행하는 것이 적절하다. Wald 통계량을 통해 모형에 유의하지 않은 변수를 확인해보았고, stepwise selection을 실행하여 이에 대한 근거를 뒷받침하였다. 또한 Hosmer and Lemeshow Goodness-of-Fit Test (적합도 검정), deviance를 통해 모형이 적합하다는 것을 확인하였으며, 해석에도 크게 무리가 없으므로 진행한 분석이 타당하다고 여겨진다.

3. 결론

3.1 분석 결과 요약

분석에서는 대학원 합격률에 영향을 미치는 변수를 찾기 위해 로지스틱 회귀분석을 사용하였다. 로지스틱 회귀분석의 장점은 오즈(Odds)의 관점에서 해석 될 수 있다는 것이다. 이를 바탕으로 대학 합격률에 유의한 변수들은 gre 점수, univ점수, lor, 학부 평점이 높을수록 불합격률이 떨어지는 것으로 나타났다. 또한 사용된 변수 중 TOEFL과 추천받은 정도를 나타내는 두 변수는 반응변수에 유의하지 않음을 확인했다.

결론적으로 사용한 설명변수 대부분이 대학원 진학에서의 합격률을 설명하는 데에 유의하고, 이에 특히 영향을 많이 끼치는 변수는 gre 점수, 대학 평가 점수이다.

3.2 분석의 장점 및 한계점 설명

이번 분석은대학원 합격 여부(반응 변수)에 어떠한 설명 변수들이 영향을 많이 끼치는지에 대해, 로지스틱 회귀분석을 통해서 쉽게 파악할 수 있다는 것이 장점이다.

이 주제를 가지고 데이터를 분석하는 이유는 많은 사람들에게 접근성이 뜰어지는 정보인 대학원 진학 및 합격에 대한 정보를 분석하고, 나아가 이 정보가 필요한 사람들에게 도움이 되기 위함이다. 본 분석에서 사용한 데이터는 주로 성적과 관련한 변수들이 대부분이였다. 따라서 성적과 관련된 요인들을 분석하여, 각 변수들에 대한 오즈를 사용하면서 어떠한 요인들이 유의하게 영향을 미치는지 알 수 있었다. 모델이 적합하고 설명력 또한 높지만, 성적이 높은 경우 대학원에 잘 진학한다는 이미 널리 알려진 결과가 결론으로 나왔다는 점이 아쉬움으로 남는다. 본 분석은 잘 되었다고 할 수 있으나, 유익한 정보가 되지는 못한다는 것이다. 또한 이번 낸 결과를 가지고 어떠한 성적 변수들이 영향을 주는지는 파악이 가능하지만 그 외 사회적, 경제적 요인과 관련한 정보를 담은 변수가 없어 보다 폭넓은 분야의 요인들 중에서, 어떤 요인이 가장 대학원 입학 합격률에 영향을 많이 끼치는지는 알 수 없다는 한계를 가지고 있다.

3.3 추가 연구사항 제안

앞서 언급한 것과 같이, 조금 더 넓은 범위의 데이터가 보강이 된다면 보다 다양한 요소를 고려한 분석을 할 수 있을 것으로 생각한다. 또한 이번 분석에서는 로지스틱 회귀모형을 이용하여 모델을 적합하였으나, 회귀를 기반으로 하는 다양한 모델들을 더 고려하여 더욱 성능이 좋은 분류 모델을 구축할 수 있을 것이다.

데이터의 한계로 인해, 이번 분석에서는 우리나라의 데이터가 아닌 인도의 데이터를 다루었다. 이는 추후 데이터가 보강된다면 자연스럽게 우리나라의 데이터에도 적용할 수 있을 것이다.

참고문헌

- 4년제 대졸자의 대학원진학 결정요인 분석 (RISS)

https://www.riss.kr/search/detail/DetailView.do?p_mat_type=1a0202e37d52c72d&control_no=148c70343477858547de9c1710b0298d

- USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

http://www3.dsi.uminho.pt/pcortez/student.pdf

- 2019 한눈에 보는 교육 주요내용

https://overseas.mofa.go.kr/oecd-ko/brd/m_20807/down.do?brd_id=20132&seq=85&data_tp=A&file_seq=1

- 대학원의 교육/연구/ 경쟁력 확보 방안

http://www.prism.go.kr/homepage/researchCommon/downloadResearchAttachFile.do;jsessionid=9F7C55A452888AC969B20877BF5A70BB.node02?work_key=001&file_type=CPR&seq_no=001&pdf_conv_yn=N&research_id=1341000-201300062