**Admission\_Predict**

**[변수설명]**

- 결측치 없음을 확인

- Serial No. 변수는 쓸모없는 변수이므로 제거하기로 결정.

#Serial No. : 대학생정보에 부여된 임의의 숫자

#GRE Score : 미국의 대학원수학자격시험 점수(0~346)

#TOEFL Score : 토플 점수(0~120)

#University Rating : 진학 대학교의 등급(1~5)

#SOP : 대학원에 진학하는 동기와 계획을 정리하는 학업계획서 점수(1~5)

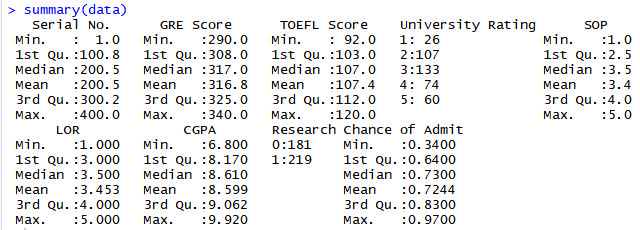
#LOR : 추천서의 점수(1~5)

#CGPA : 학점(0~10)

#Research : 실험 경험 유무(0 or 1)

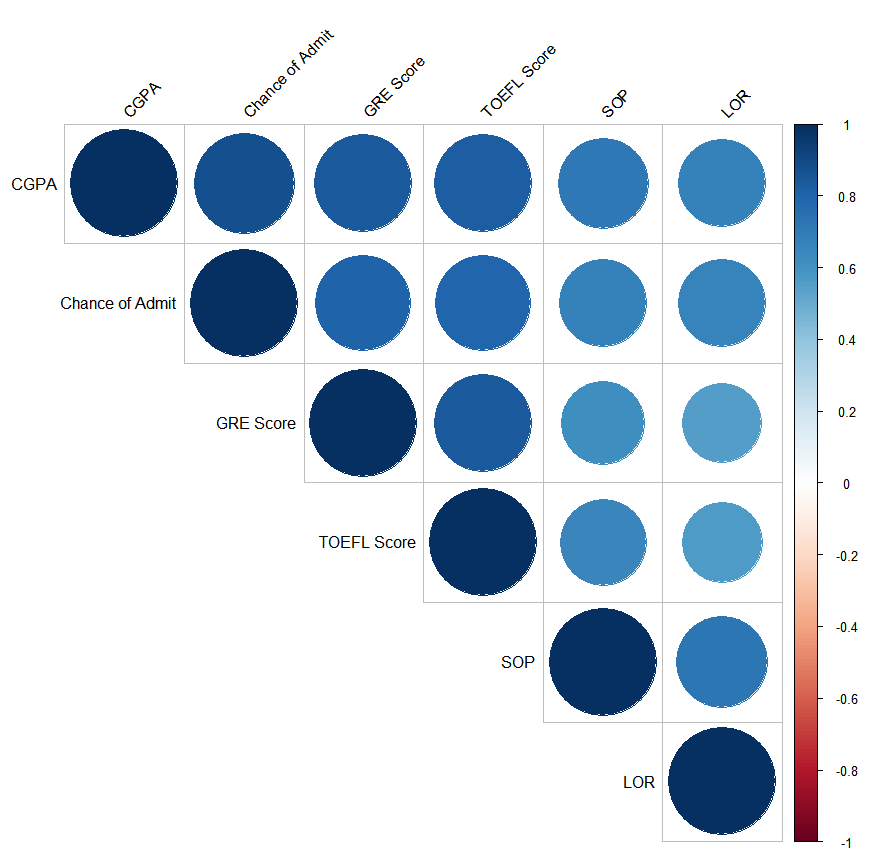
#Chance of Admit : 대학원에 합격할 확률. (0~1)

└ 대학원에 있는 사람 중 데이터의 한 행과 입학데이터가 같은 사람의 비율로 보았음.



**EDA**

**[연속형 변수간 corrplot그리기]**

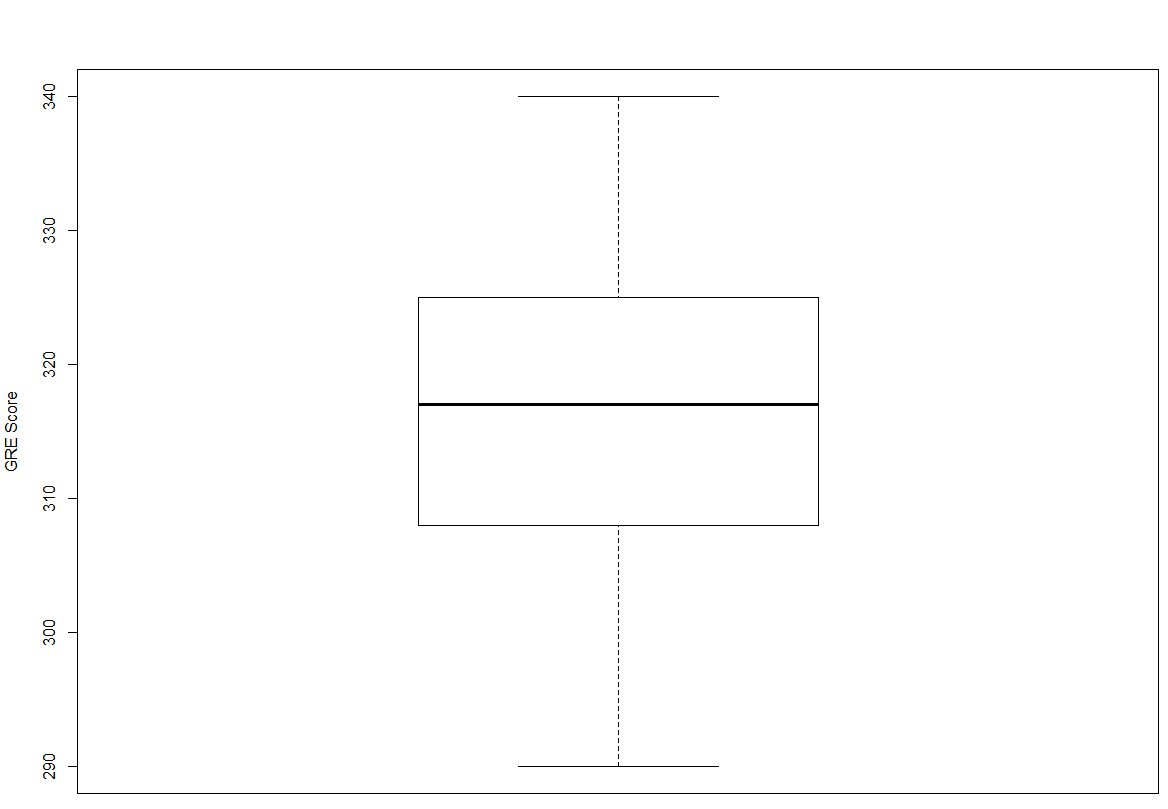
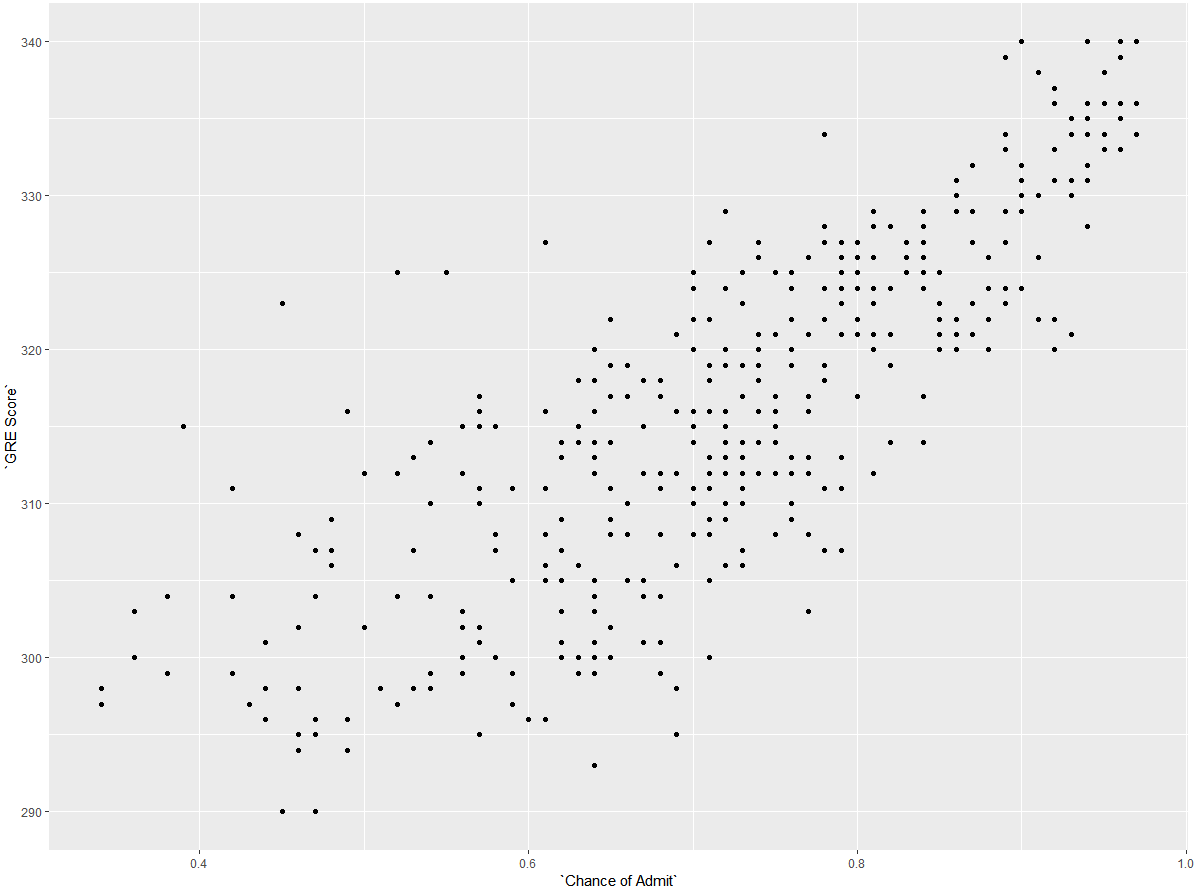


→ 대부분의 변수들 간의 선형성이 매우 크다고 볼 수 있다.

따라서 만약 회귀분석을 진행한다면 다중공선성이 의심 될 수 도 있다.

**EDA - 목적변수와의 관계**

**[GRE Score]**

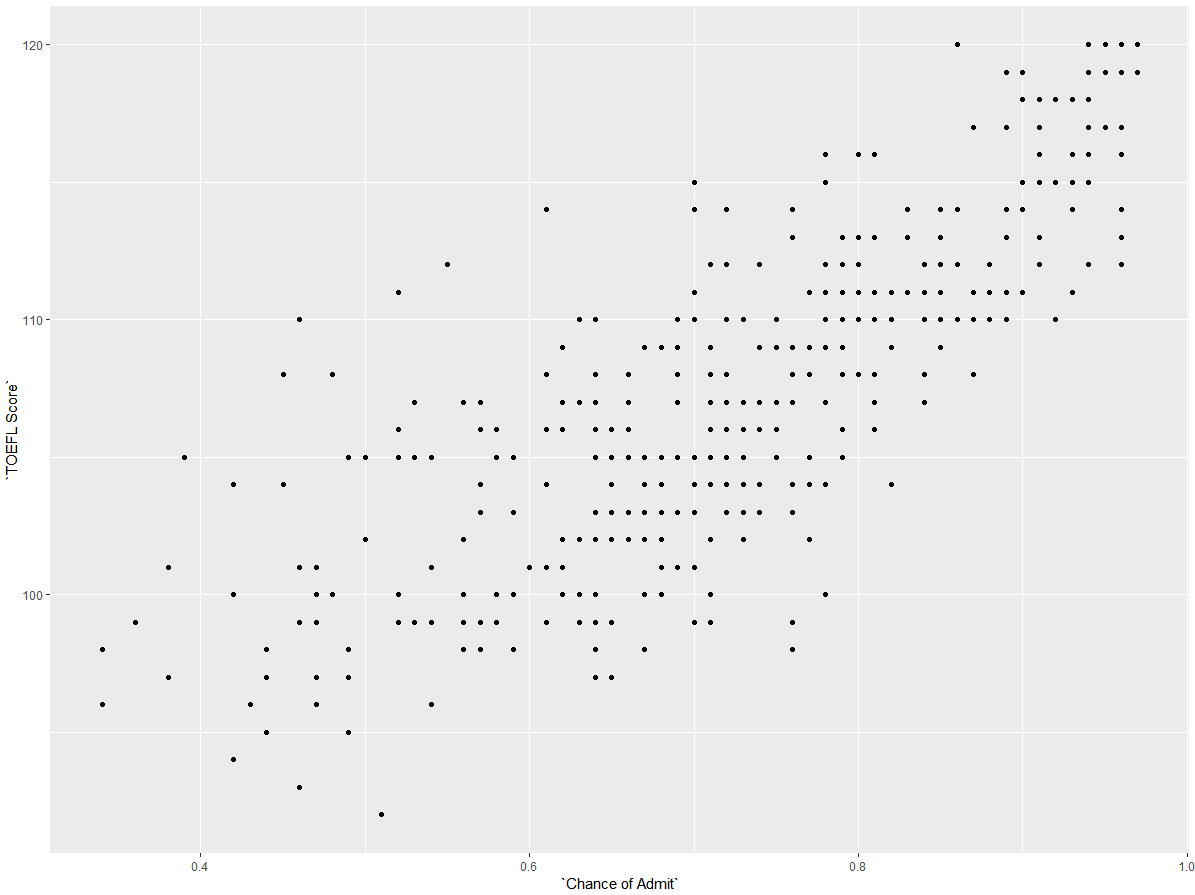
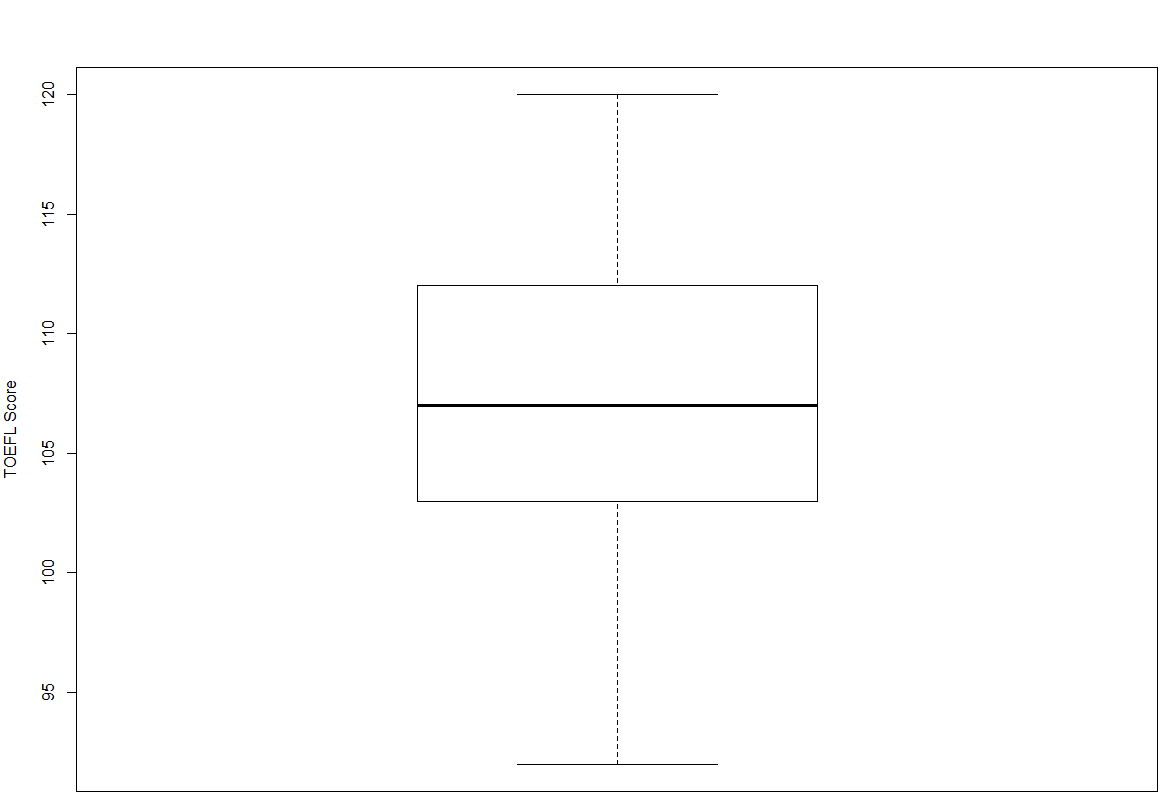
GRE점수가 높아질수록 대학원의 합격률이 높아진다고 볼 수 있다.

<Boxplot>

거의 대칭적인 분포모양을 하고 있다.

**EDA - 목적변수와의 관계**

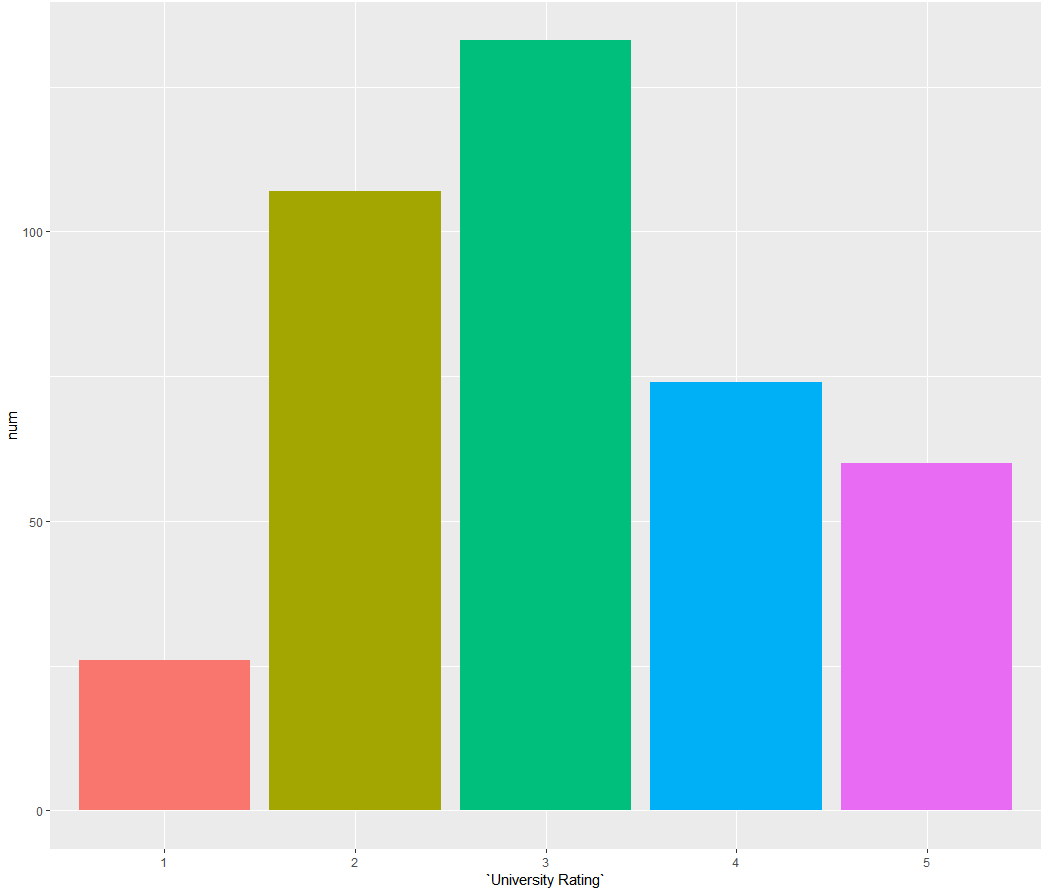
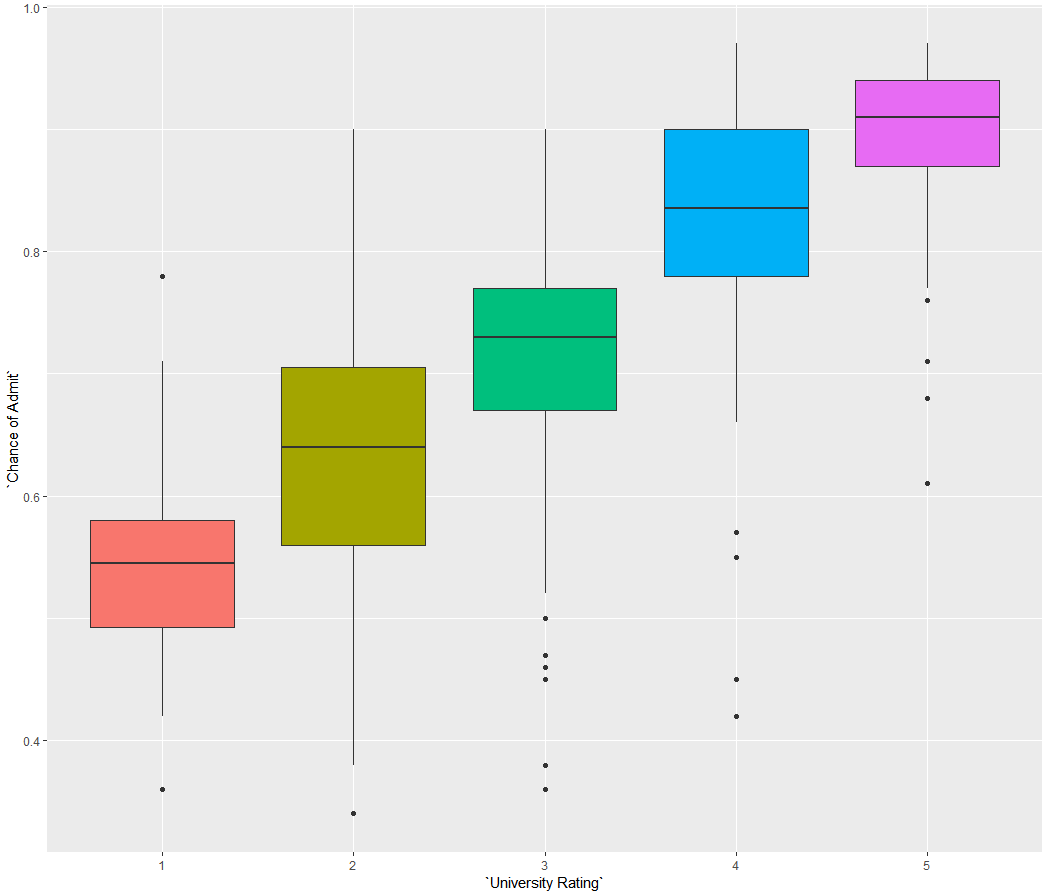
**[TOEFL Score]**



토플점수가 높아질수록 대학원의 합격률이 높아진다고 볼 수 있다.

살짝 오른쪽으로 치우쳐져 있는 듯 하나 무시할 수 있는 정도라고 판단.

**[University Rating]**

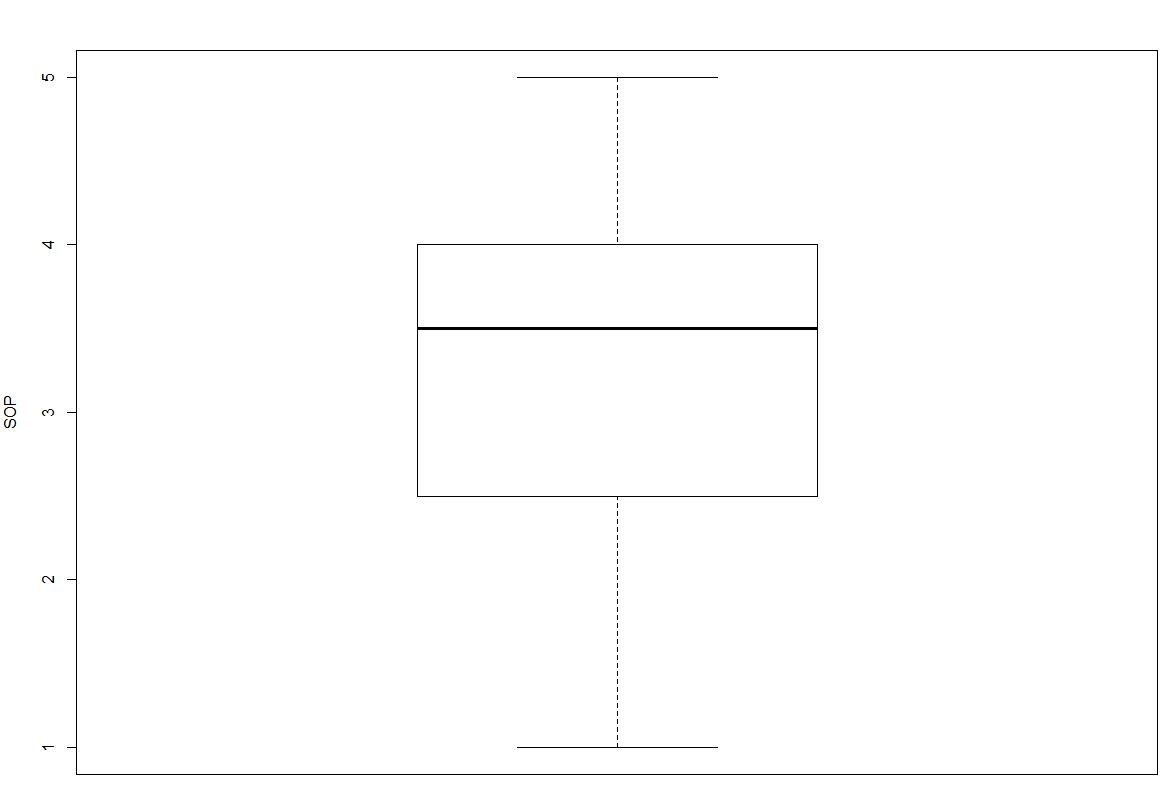
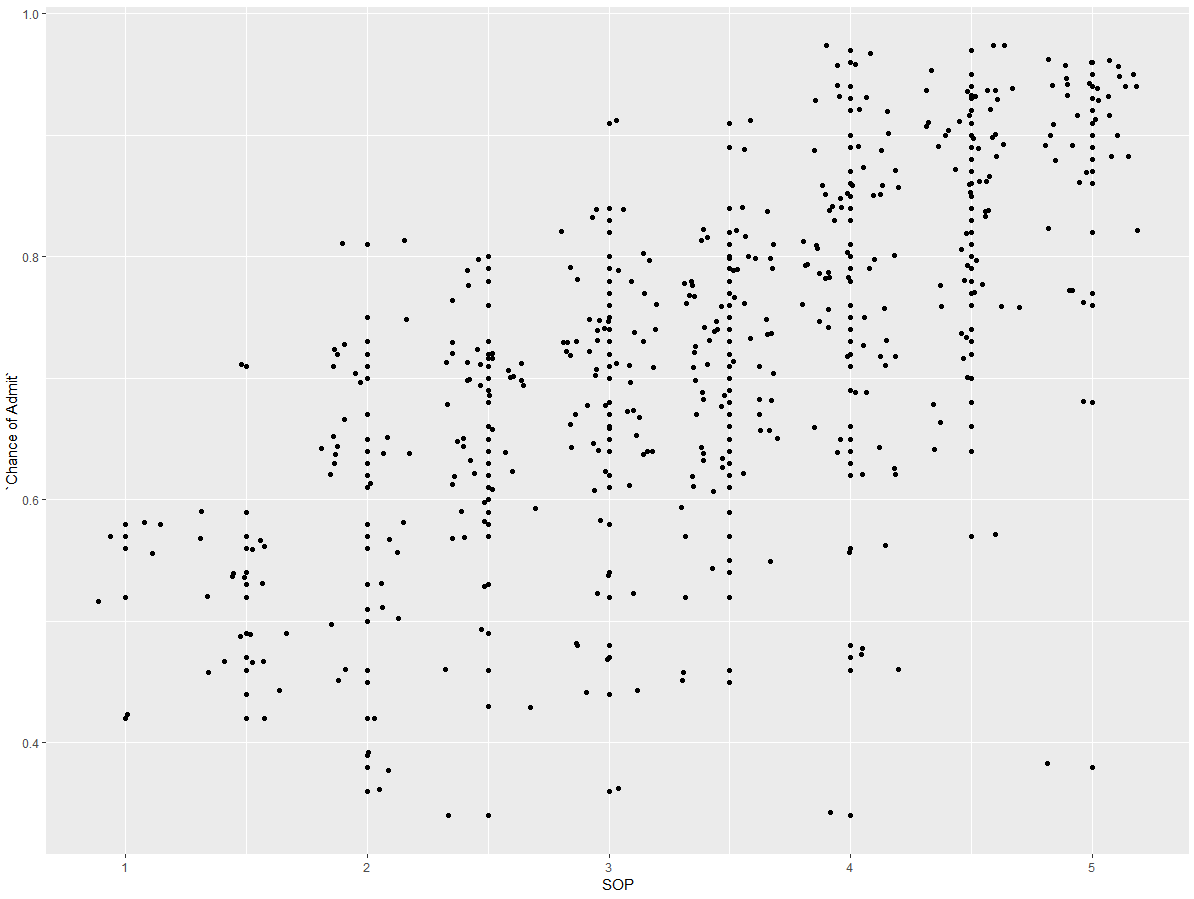
 ****

관측치의 개수로 보면 3>2>4>5>1 순으로 관측치의 갯수가 크다.

대학의 등급이 높을수록 대학원의 합격률이 높아진다고 볼 수 있다.

**EDA - 목적변수와의 관계**

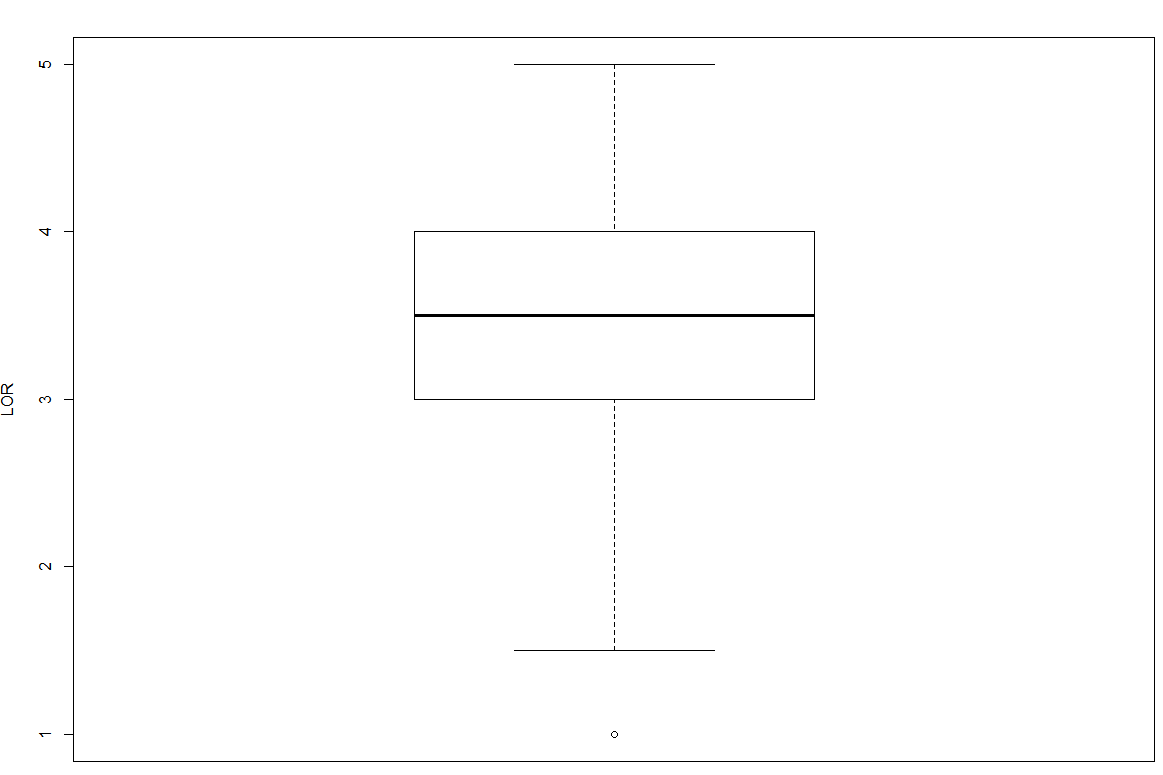
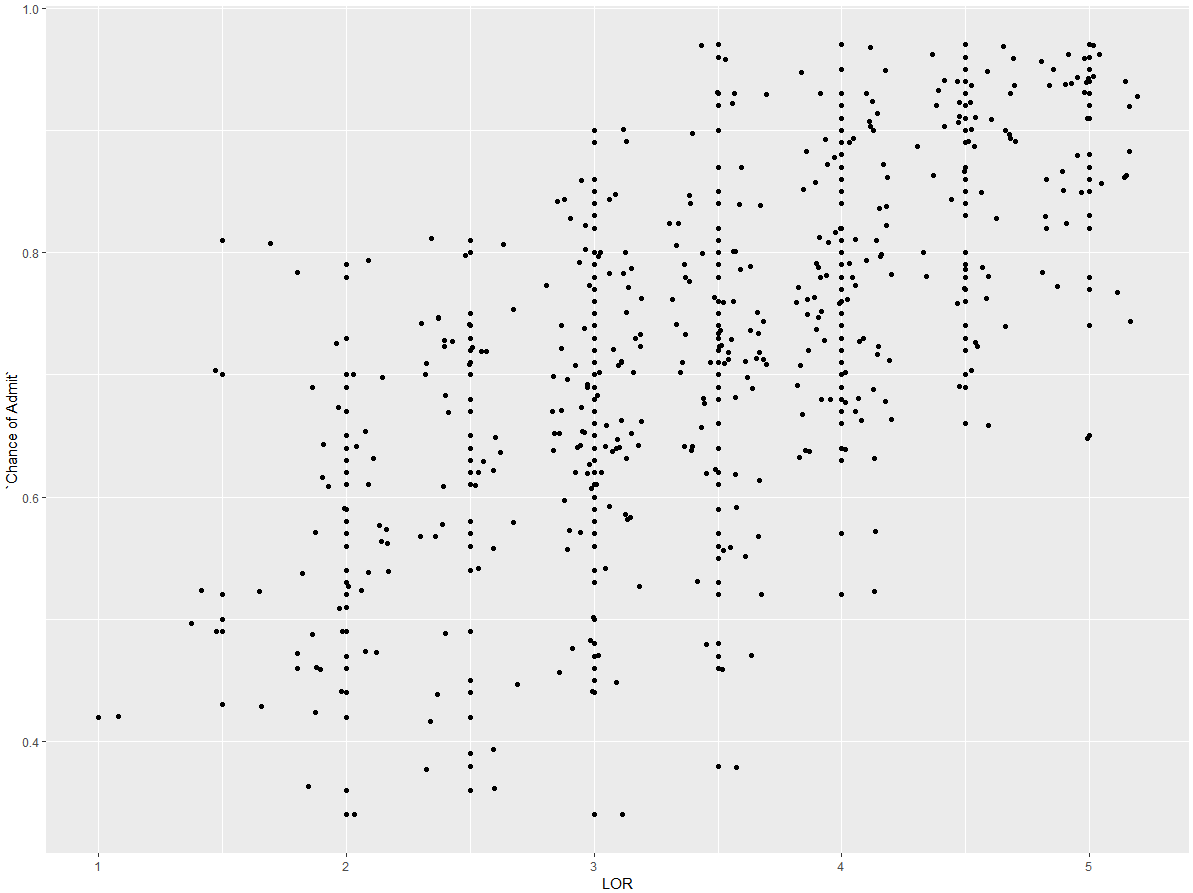
**[SOP]**

학업계획서의 점수가 높아질수록 대학원 합격률이 높아진다고 볼 수 있다..

오른쪽으로 치우쳐진 분포모형임을 알 수 있다.

**[LOR]**

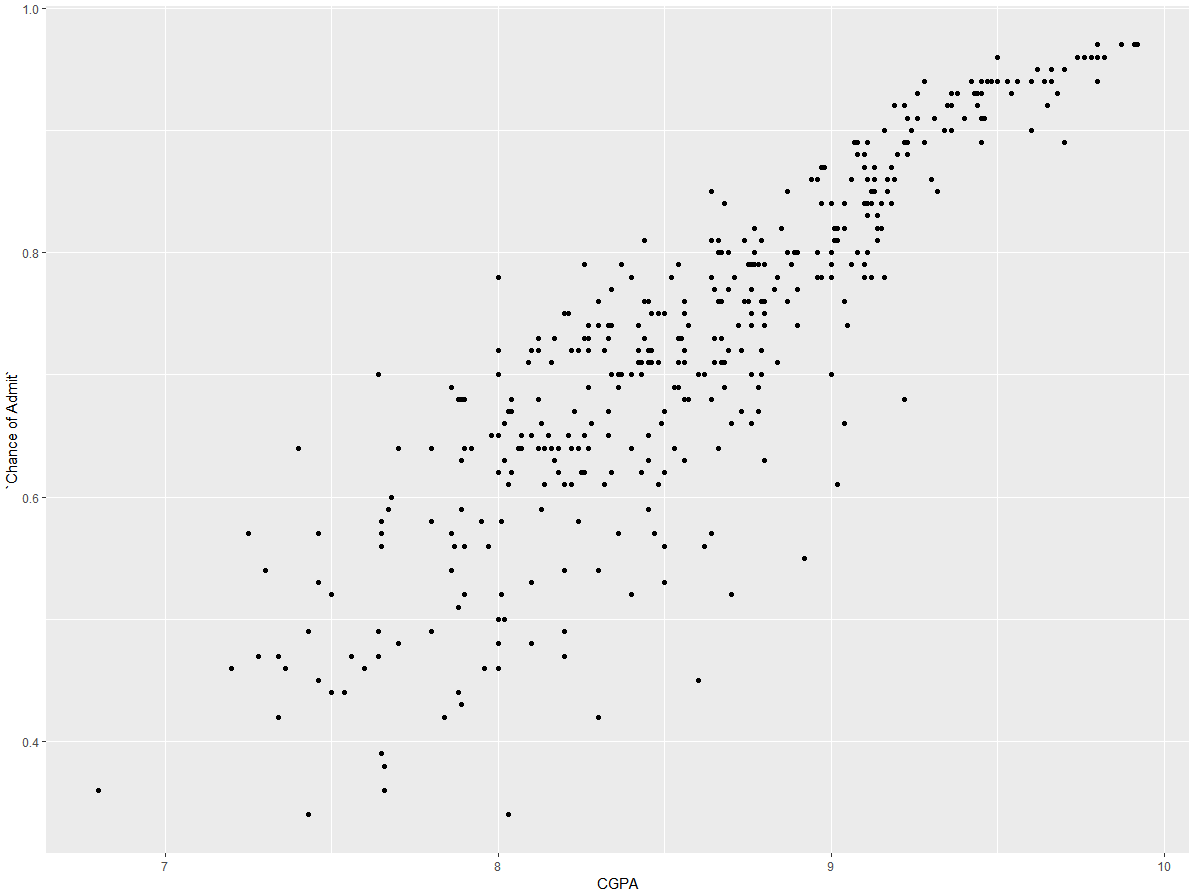
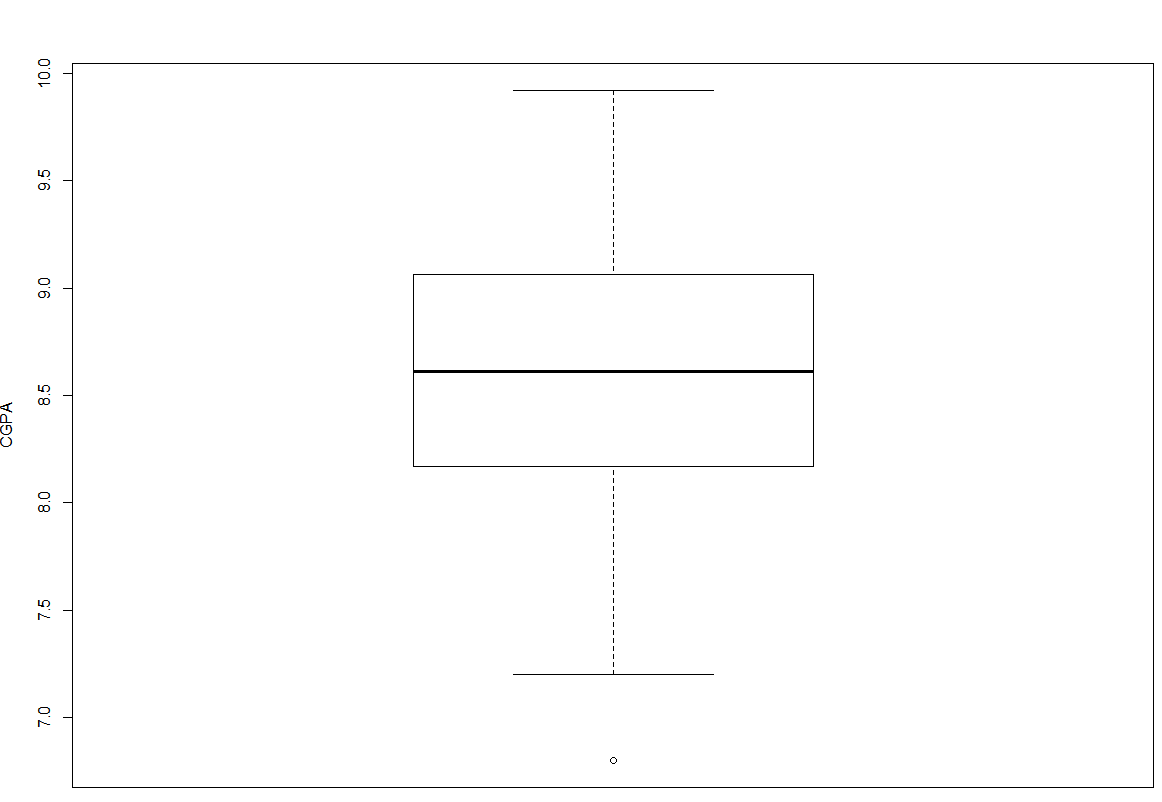
추천서의 점수가 높아질수록 대학원의 합격률이 높아진다고 볼 수 있다..

분포모형이 대칭적이다.

이상치 1이 발견되지만 방해되지 않아서 제거하지 않음.

**EDA - 목적변수와의 관계**

**[CGPA]**

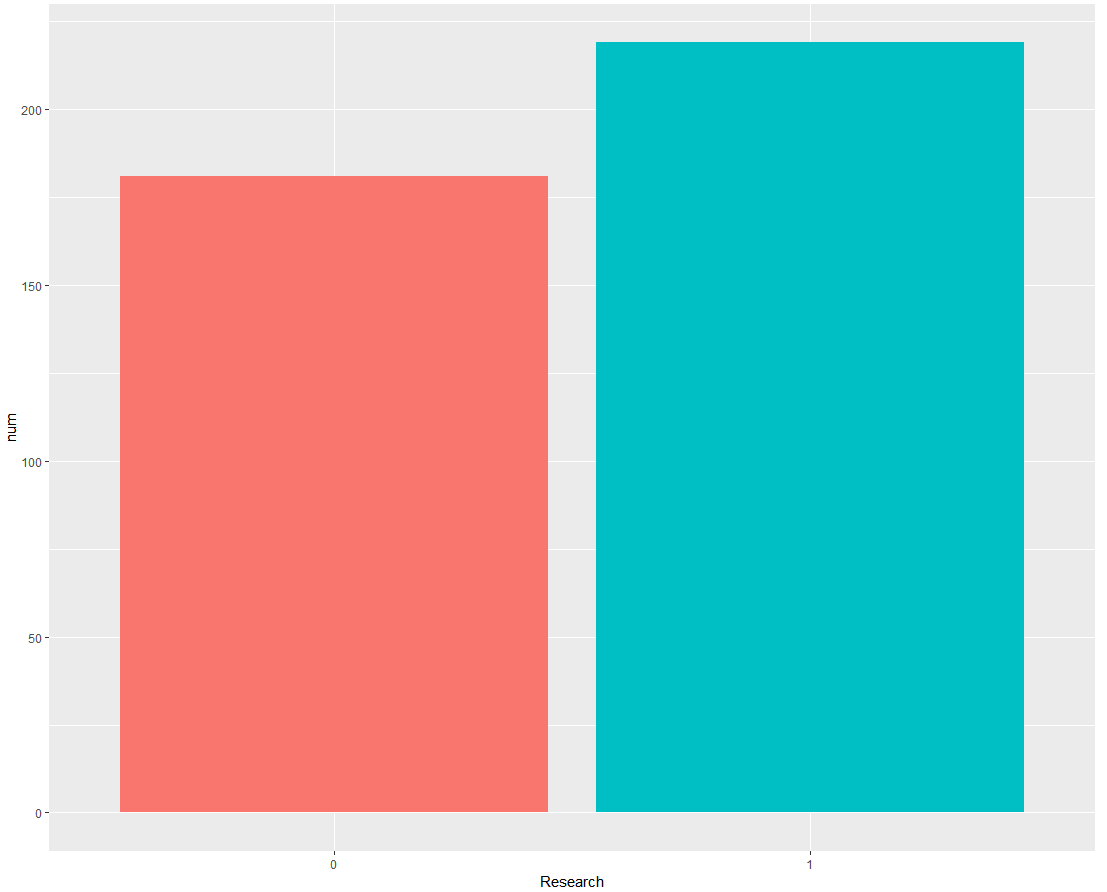
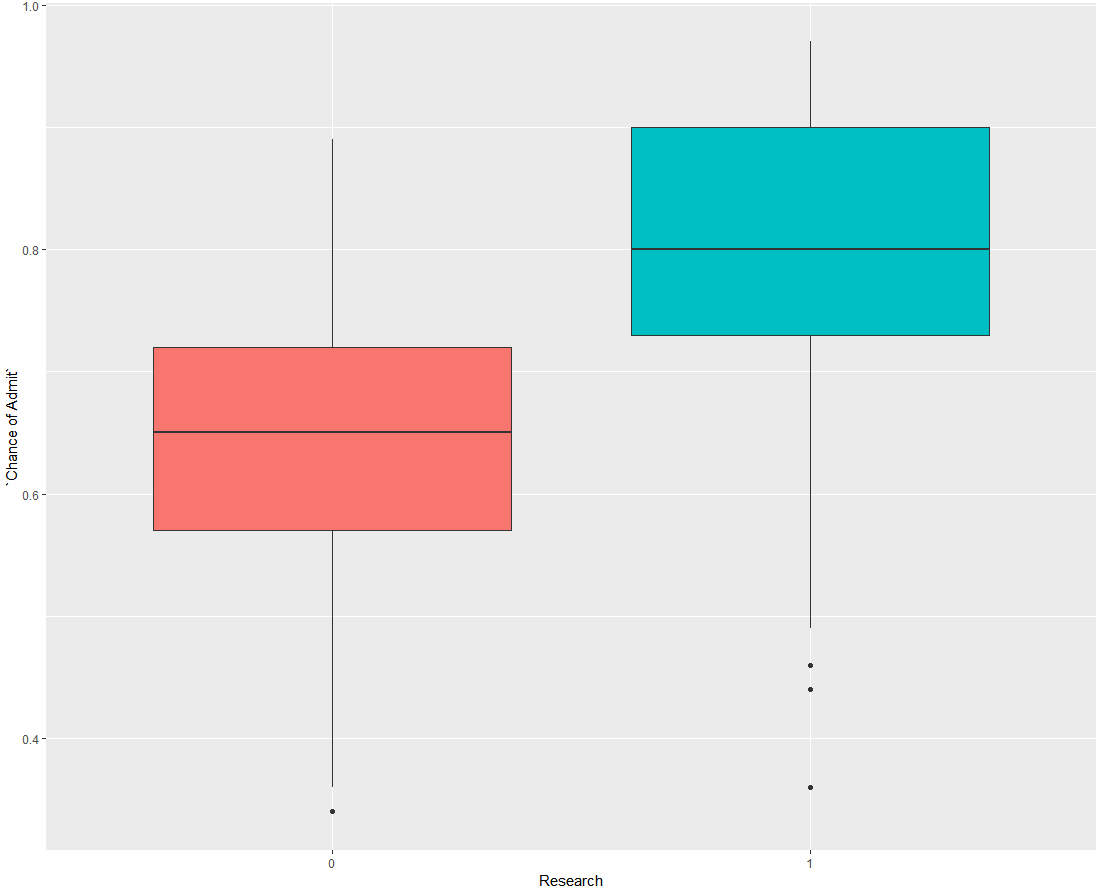


학점이 높아질수록 대학원의 합격률이 높아진다고 볼 수 있다..

분포모형이 대칭적이다.

이상치 1이 발견되지만 방해되지 않아서 제거하지 않음.

**[Research]**

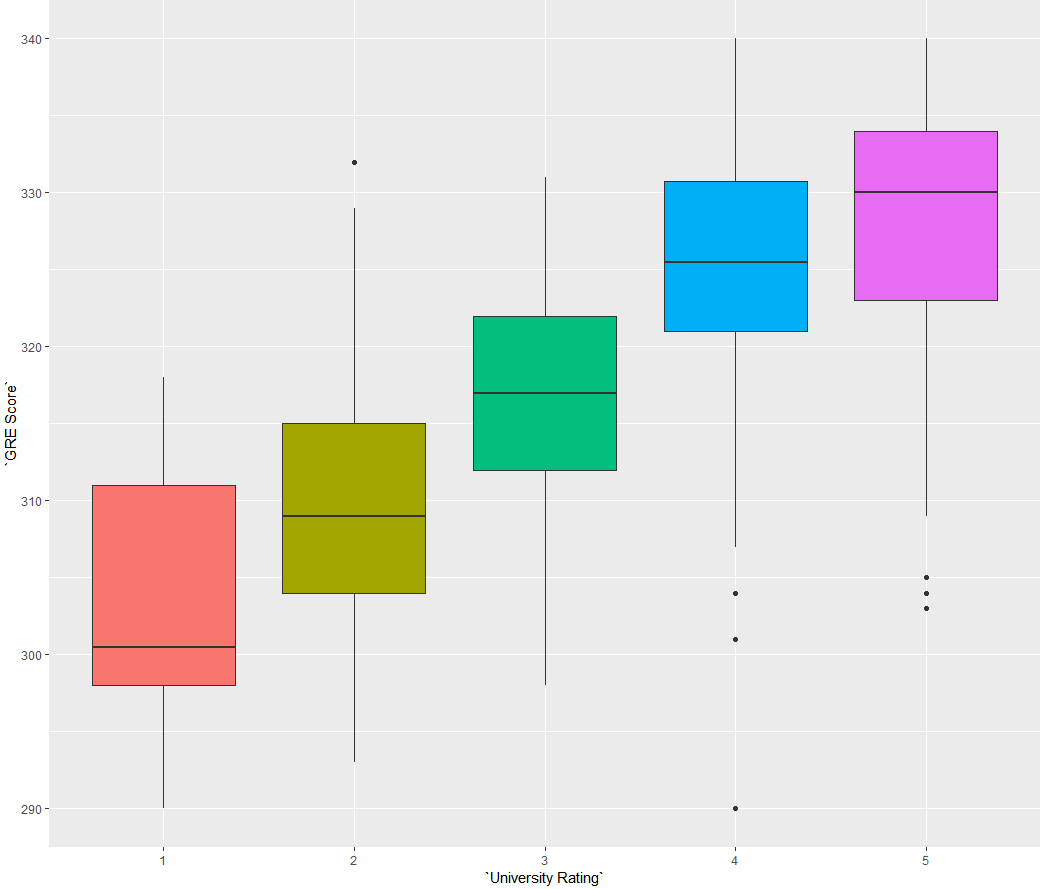
 

실험경험이 있으면 대학원의 합격률이 더 높다고 볼 수 있다.

실험경험이 있는 사람이 더 많다.

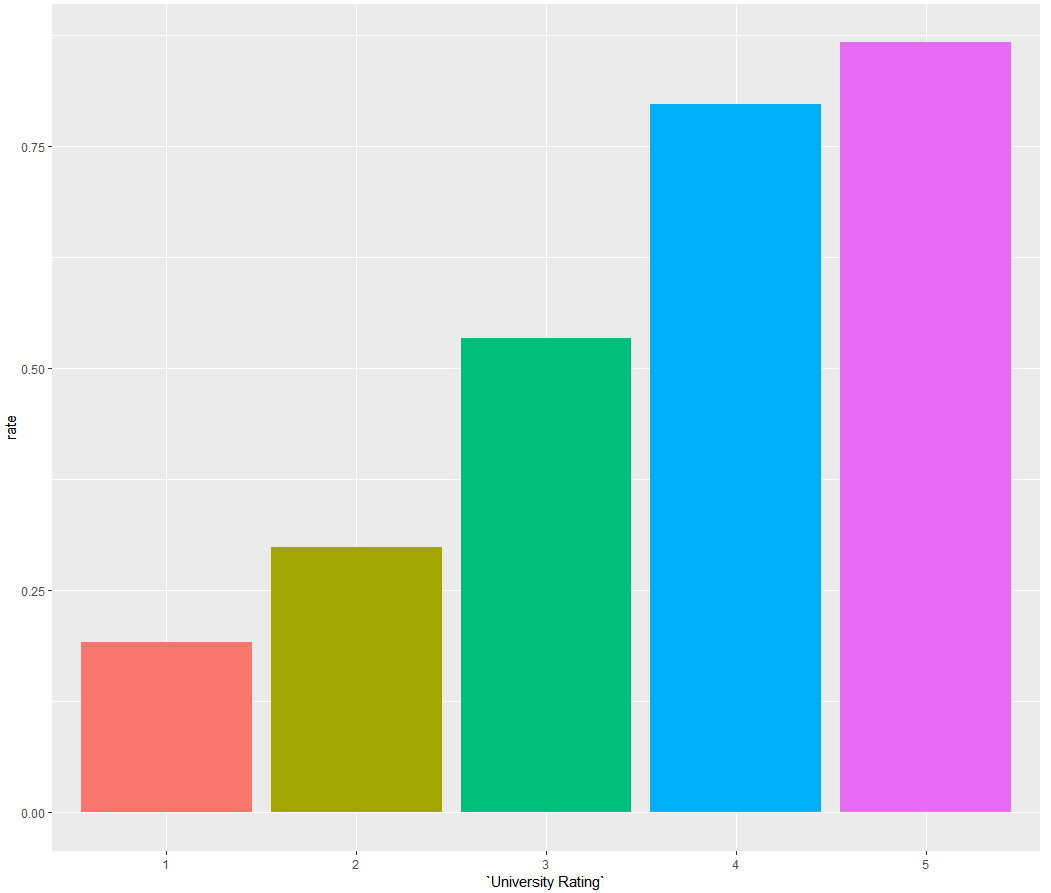
**EDA - 반응변수들 간의 관계**

**[University Rating - GREScore]**



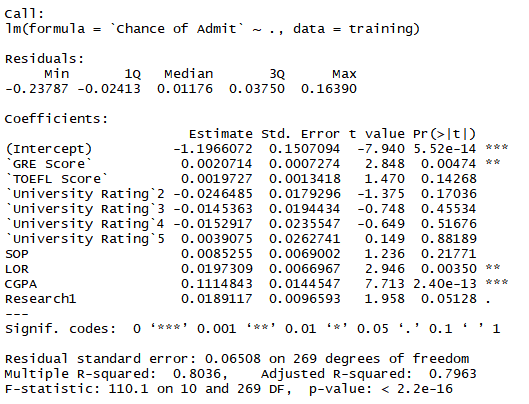
대학등급이 높을수록 점수또한 높다고 볼 수 있다.

**[University Rating – Research]**



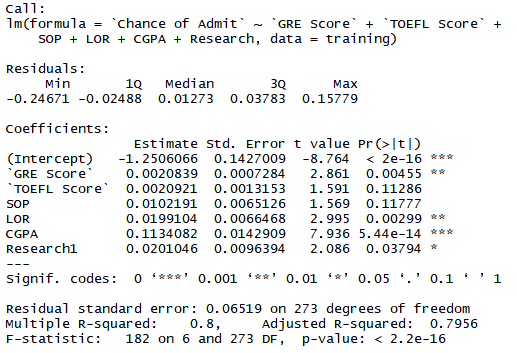
대학등급이 높을수록 실험경험이 있을 확률이 높다.

**회귀분석**



모델은 적합하나 유의하지 않은 변수들이 많음.

→단계별 회귀분석 실행



Validation set 을 예측해보았을 때 약 80%의 적합도를 보이고

RMSE를 구해보았을 때 약 0.06214값을 지님.

vif계수가 10을 넘지않으므로 다중공선성을 의심하지 않아도 된다고 생각.

대학등급 변수만 제거됨을 알 수 있다.

R-Square 값은 0.7956으로 약 80%의 적합도를 보여준다.

**LASSO & RIDGE**

#리지는 변수의 계수를 줄여준다. 다중공선성을 막는다.

#라쏘 계수를 줄여주는데 줄이다가 계수가 0이 되면 변수가 사라지므로 변수선택의 의미도 가진다. 따라서 많은 변수를 다룰 때 사용된다.

#알파값으로 혼합한 것이 엘라스틱

xx <- model.matrix(`Chance of Admit`~.-1,data=data)

* factor 형 변수를 dummy 화 시켜준다.

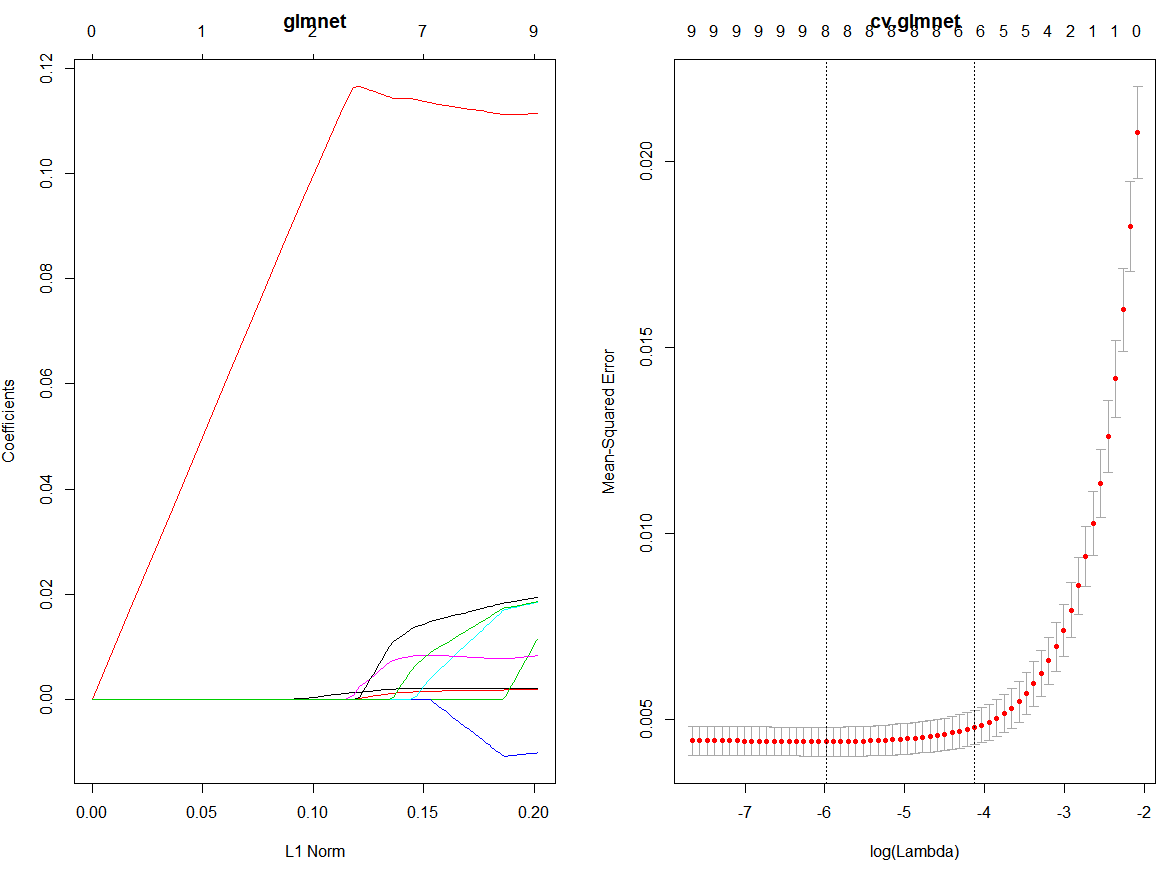
$lambda.min

[1] 0.00251610

그래프의 최솟값, 교차검증의 오차를 최소로 하는 람다값

$lambda.1se

람다 값 중 1표준오차만큼 크지만 훨씬 단순한 모델을 만드는 람다값



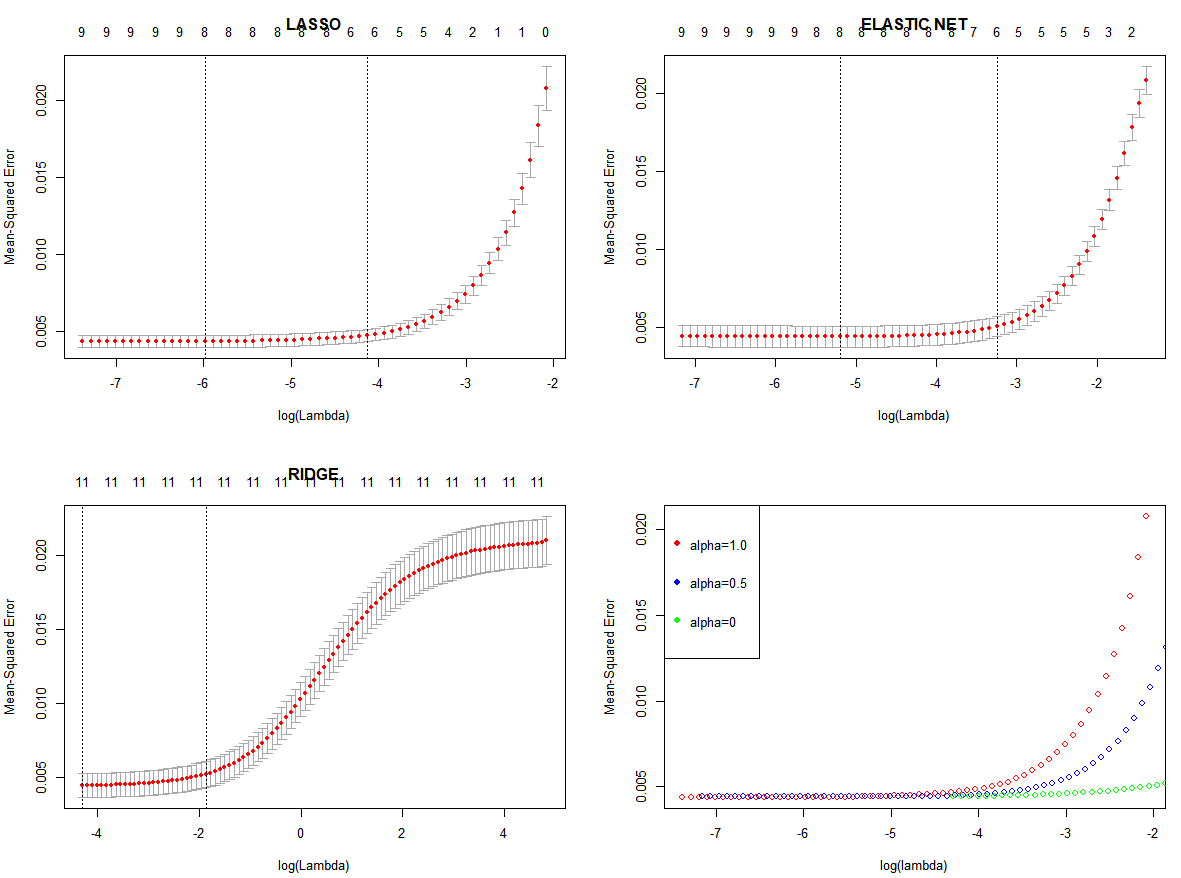
lambda가 증가함에 따라 계수가 어떻게 변하는지 보여주는 그래프

lambda가 증가함에 따라 표준오차가 어떻게 변하는지 보여주는 그래프

**LASSO & RIDGE**



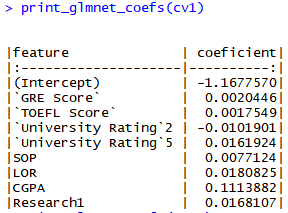
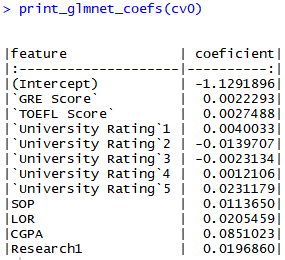
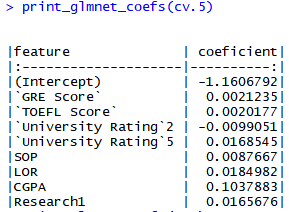
* lambda. 1se을 사용했을 때 변수의 개수를 1개 줄여준다.



(여기서 ELASTIC 은 알파값이 0.5일떄의 모형)

리지 모델이 가장 안정적인 것을 알 수 있다..

LASSO 모델 ELASTIC모델 RIDGE모델

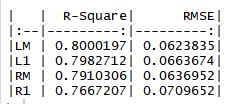
변수의 개수가 많지 않으므로 LASSO 모델은 좋지 않다고 생각.

그리고 현재 일반회귀를 돌릴 시 다중공선성이 의심되므로 RIDGE 모형을 선택.

변수의 개수가 많지도 않아서 모델의 단순화 보다는 교차검증의 오차를 줄이는 데에 의미가 있다 생각하여 람다를 lambda.min으로 사용..

**LASSO & RIDGE**

**[LASSO & RIDGE 비교]**



라쏘방법과 표준오차를 lambda.min으로 한 모델이 RMSE가 가장 작고 R-square또한 가장 높으므로 가장 좋은 모델임을 알 수 있다.

**[ELASTIC NET]**

**표준오차를 lambda.1se로 잡았을 때**

**Alpha=0.96 rmse=0.06533273**

**표준오차를 lambda.min로 잡았을 때**

**Alpha=0.92 rmse=0.06234282**

> A와 B의 최적의 alpha값을 넣어 만든 모델의 rmse를 비교해보면 표준오차를 lambda.min으로 하는 모델이 가장 좋다고 볼 수 있다.

**최종적으로 라쏘 리지에서 가장 좋은 모델이었던 라쏘모형에서 표준오차를 lamda.min으로 잡은 모델과 비교해보면**

**LM[2] ; A[which.min(A)]**

**[1] 0.06238354**

**[1] 0.06234282**

**ELASTIC NET에서 표준오차를 lambda.min로 잡았을 때가 근소한 차이로 가장 좋다.**

**RANDOM FOREST**

**전체 변수를 모두 넣고 모형 만들었을 때 rmse=0.0729884**

**선형회귀에서 유의하다고 판단되는 변수(GRE Score,TOEFL Score,SOP,LOR,CGPA,Research)만 넣고 모형 만들었을 때**

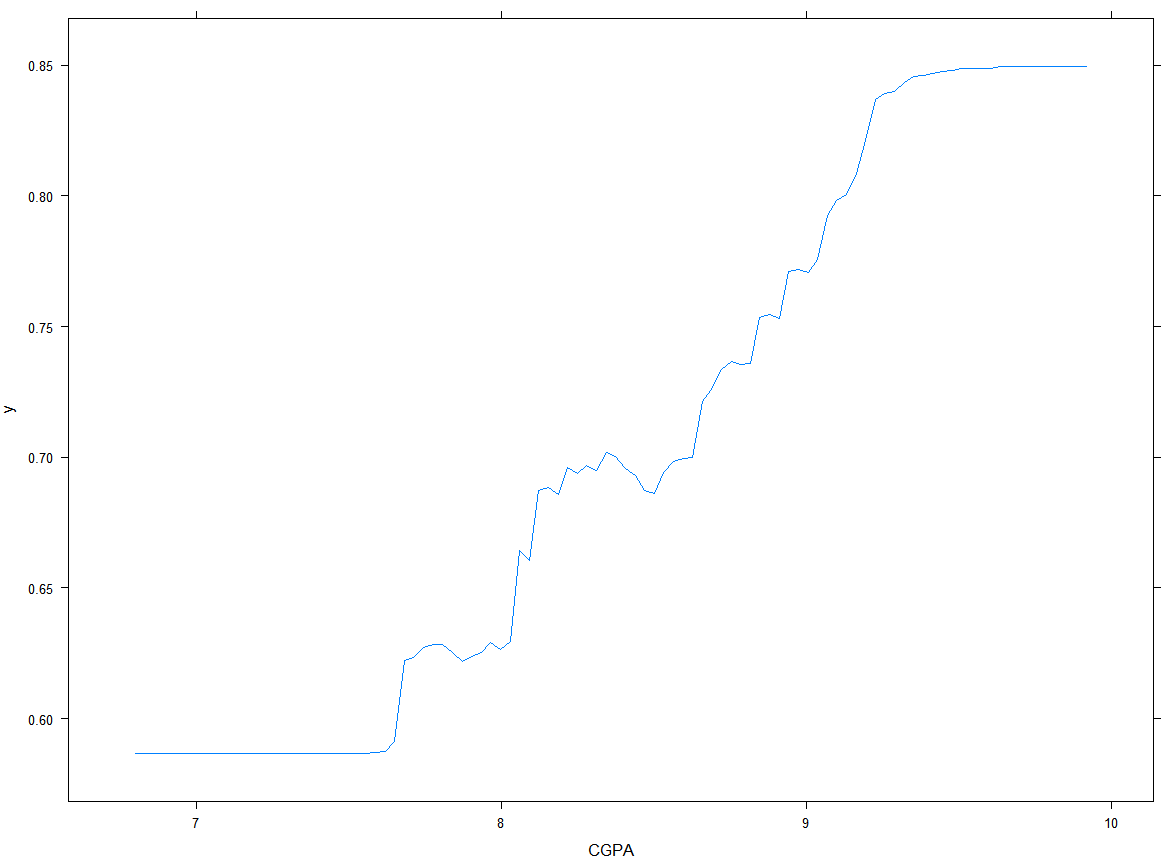
**rmse= 0.0729884**

**같게 나오는 이유? 나머지 변수들은 영향을 안 미치는 있으나 없으나한 변수들이었다.**

**아직 표준오차를 lambda.min로 잡았을 때 ELASTIC NET 모형이 rmse관점에서 가장 좋다.**

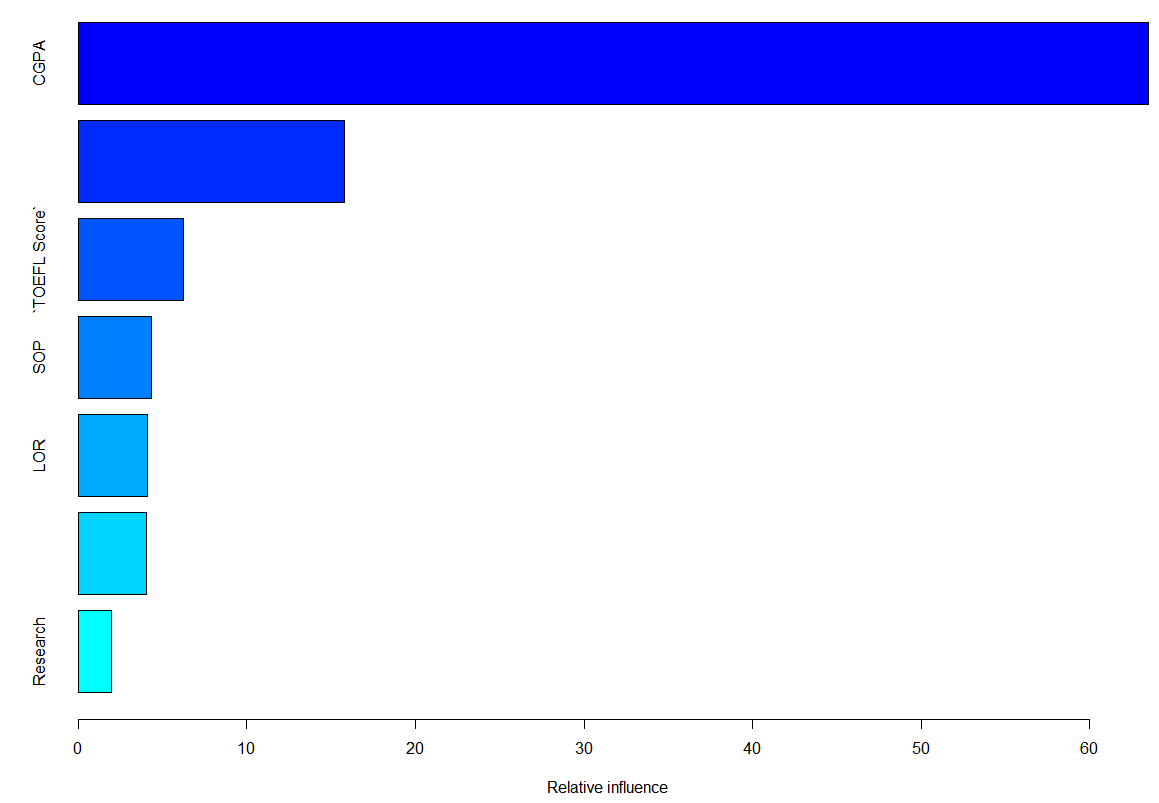
**GBM**

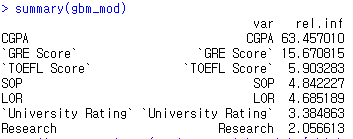
**(tree 개수 100일 때)**

CGPA변수가 가장 중요한 변수

양의 상관관계를 가진다.

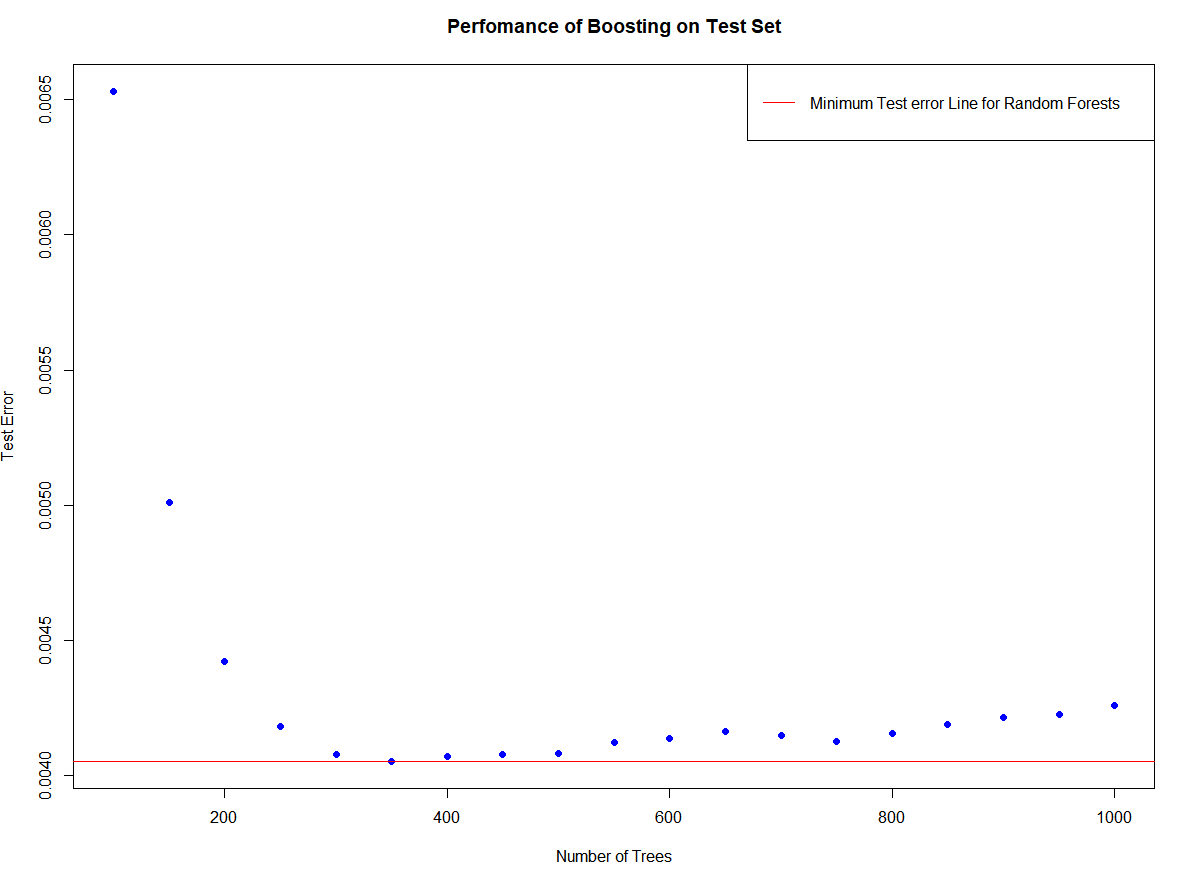
두번째로 중요한 변수는 GRE Score





**GBM**

**[최적의 나무 갯수 확인하기(100에서 1000까지 50씩 증가시킴)]**



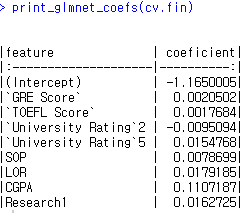
* 나무의 갯수가 350일때 가장 적은 에러를 가진다고 볼 수 있다.

Tree=350일 때 예측값을 구해서 rmse 를 구해보면 0.0636606 가 나온다.

**ELASTIC NET에서 표준오차를 lambda.min로 잡았을 때**

**rmse 관점에서 가장 좋은 모델!!!**

**결론**



모델링을 다 해본 후 가장 좋은 모델의 선형식을 살펴보면 계수들 중 CGPA의 계수가 가장 큰 것을 볼 수 있습니다. 즉, 학점이 가장 대학원 합격률에 영향이 크다고 볼 수 있습니다.