Multivariate Data Analysis
Assignment #1 – Multivariate Linear Regression (MLR)

공과대학
산업경영공학부
2015170867
이종현

[Q1] https://www.kaggle.com/mohansacharya/graduate-admissions
1 차 과제로 석사과정 합격여부를 판단하는 데이터셋을 모델링하였다. 선정한 이유는 400 개의 풍부한 데이터셋과 7 개의 설명변수를 가지고 있고, 정량적 사고를 사용하지 않더라도 설명변수들이 종속변수에 영향을 끼친다는 것을 직감적으로 알 수 있기 때문이다.

[Q2]
설명변수:
GRE Scores (Graduate Record Examinations)
TOEFL Scores
University rating
SOP strength (Statement of Purpose)
LOR strength (Letter of Recommendation)
CGPA (Cumulative Grade Point Average)
Research Experience

종속변수:
Chance of admittance

1. 이 데이터는 종속변수와 설명변수들 사이에 선형 관계가 있다고 가정할 수 있다 . 대학 입시가 그렇듯, 석사과정 합격 여부 또한 한 학생이 학부생활동안 얼만큼의 좋은 성적을 얻었는지(GRE, TOEFL, CGPA)에 따라 나뉩니다. 더 좋은 대학출신(University rating)이면 합격할 확률 또한 증가하고, 자소서와 추천서(SOP, LOR)의 신뢰성 또한 합격여부를 나눈다. 연구경험 (Research Experience)이 있으면 분명 합격할 확률 또한 증가할것이다.

2. 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 성적들(GRE, TOEFL, CGPA)이다. 실제로 학생들이 입시를 할때 합격 여부가 제일 많이 나뉘는 부분이 정량적으로 수치화되는 성적이라고 판단하기 때문이다.

3. 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수는 “Serial No.” 이다. 학생들의 고유번호를 선정하는 설명변수이기 때문에, 그 학생의 석사과정 합격여부엔 영향이 없을거라고 판단할 수 있다.

리포트를 진행하기에 앞서, 위 링크에서 다운받은 “Admission_Predict.csv” 파일의 마지막 행인 ‘Chance.of.Admit’을 2 번째 행으로 옮겼다고 명시한다.

[Q3]

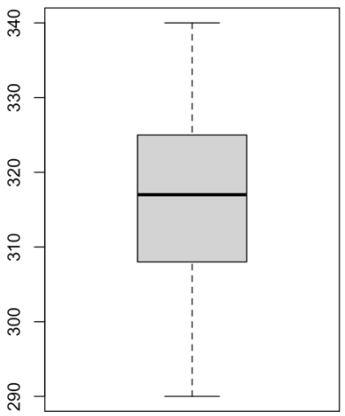
R code

```
admission <- read.csv("Admission_Predict.csv")
nStud <- nrow(admission)
nVar <- ncol(admission)
id_idx <- c(1)

admission_data <- cbind(admission[,-c(id_idx)])
describe(admission_data)

scale_admission <- scale(admission_data)
boxplot(scale_admission)
boxplot(admission_data$GRE.Score,xlab = "GRE")
boxplot(admission_data$TOEFL.Score,xlab = "TOEFL")
boxplot(admission_data$University.Rating,xlab = "rating")
boxplot(admission_data$SOP,xlab = "SOP")
boxplot(admission_data$LOR,xlab = "LOR")
boxplot(admission_data$CGPA,xlab = "CGPA")
boxplot(admission_data$Research,xlab = "Research")
```

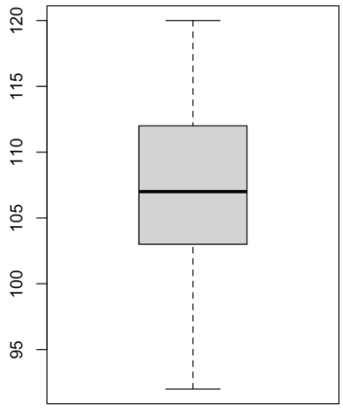
Output

```
> describe(admission_data)
                  vars   n   mean    sd median trimmed   mad    min    max range  skew kurtosis   se
Chance.of.Admit      1 400   0.72  0.14   0.73    0.73  0.13   0.34   0.97  0.63 -0.35    -0.41 0.01
GRE.Score            2 400 316.81 11.47 317.00  316.85 11.86 290.00 340.00 50.00 -0.06    -0.72 0.57
TOEFL.Score          3 400 107.41  6.07 107.00  107.33  5.93  92.00 120.00 28.00  0.06    -0.60 0.30
University.Rating    4 400   3.09  1.14   3.00    3.07  1.48   1.00   5.00  4.00  0.17    -0.81 0.06
SOP                  5 400   3.40  1.01   3.50    3.43  0.74   1.00   5.00  4.00 -0.27    -0.69 0.05
LOR                  6 400   3.45  0.90   3.50    3.46  0.74   1.00   5.00  4.00 -0.11    -0.68 0.04
CGPA                 7 400   8.60  0.60   8.61    8.60  0.67   6.80   9.92  3.12 -0.07    -0.48 0.03
Research             8 400   0.55  0.50   1.00    0.56  0.00   0.00   1.00  1.00 -0.19    -1.97 0.02
```

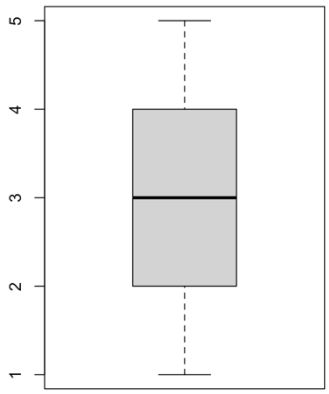
각 변수에 대한 box plot



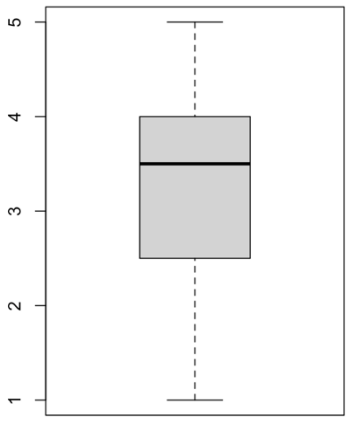
GRE



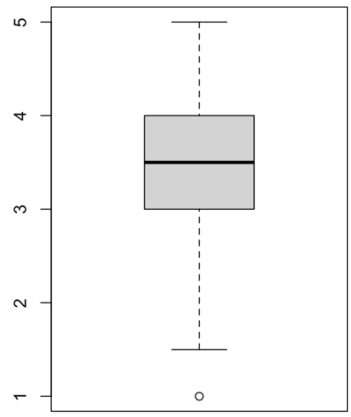
TOEFL



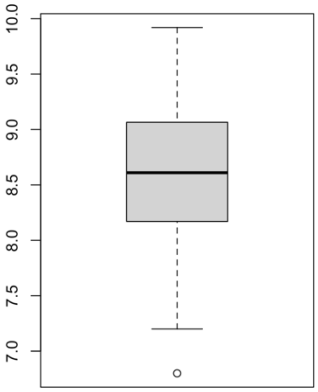
rating



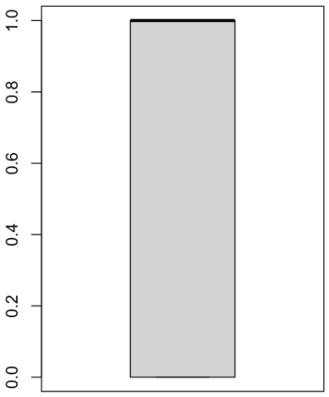
SOP



LOR

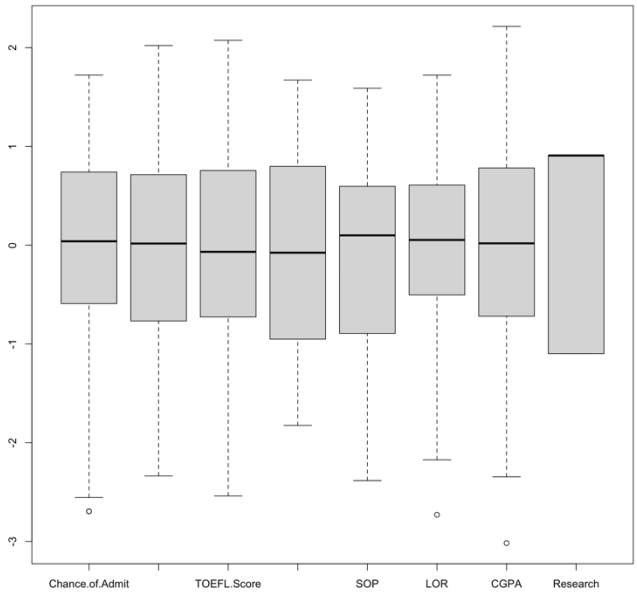


CGPA



Research

전체 변수에 대한 box plot



각 변수들에 대해 skewness 가 0, kurtosis 가 1 일때 완전한 정규분포다.
이 데이터셋에선 skewness 의 절대값이 0 과 가깝고 kurtosis 의 절대값이 1 에 가까울때 정규분포를 따른다고 가정하면, “Research” 변수의 kurtosis 값이 -1.97 이기 때문에 정규분포를 따르지 않는다.

[Q4]
Box plot 을 관찰해 보면 각 변수들에 대한 이상치가 없다. 만약 이상치가 많이 관측되는 데이터였다면, R coding 을 통해 1 분위수와 3 분위수를 기준으로 data cutoff 를 진행하였을 것이다.

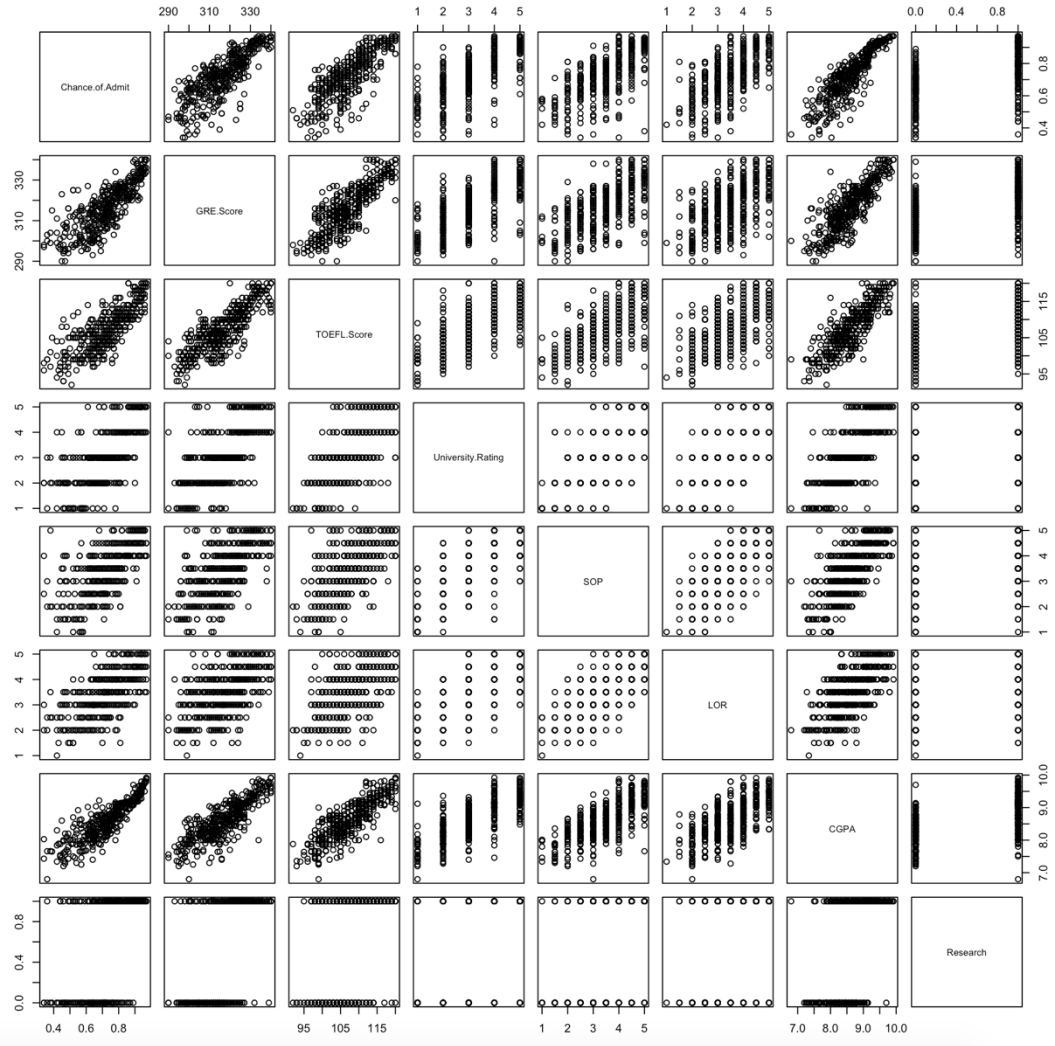
[Q5]

R code

```
pairs(~Chance.of.Admit+GRE.Score+TOEFL.Score+University.Rating+SOP+LOR+CGPA+Research,data=admission_data)

AdmCor <- cor(admission_data)
AdmCor
corrplot(AdmCor,method = "number")
```

Output

Scatter plot



Correlation plot

Correlation plot 에 따르면 CGPA 와 Chance of Admittance 의 상관계수는 0.87 로 제일 높지만, Chance of Admittance 는 종속변수이므로 제외한다. TOEFL 과 GRE 의 상관계수가 0.84 로 그 다음으로 높으므로, TOEFL 과 GRE 두 조합의 변수들이 가장 강한 상관관계를 나타낸다.

[Q6]

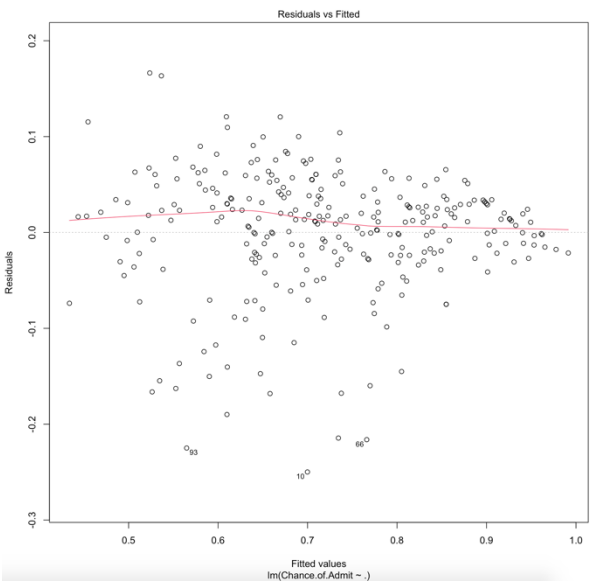
R code

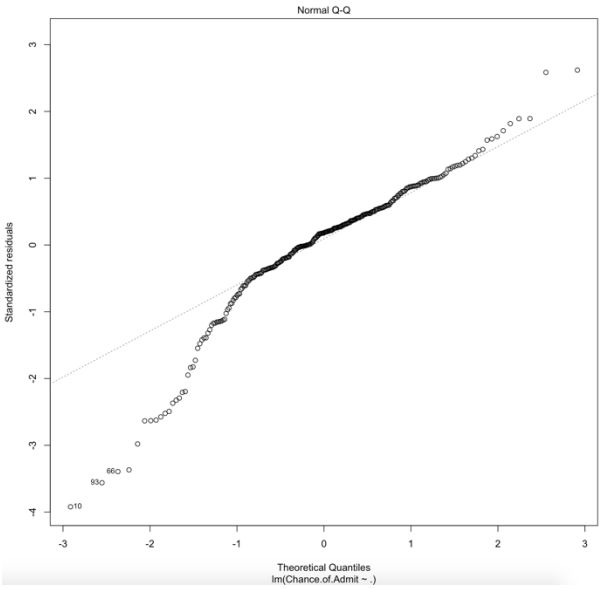
```
admission_mlr_data <- cbind(admission[,-c(id_idx)])
set.seed(12345)
admission_trn_idx <- sample(1:nStud, round(0.7*nStud))
admission_trn_data <- admission_mlr_data[admission_trn_idx,]
admission_val_data <- admission_mlr_data[-admission_trn_idx,]

mlr_admission <- lm(Chance.of.Admit ~ ., data = admission_trn_data)
mlr_admission
summary(mlr_admission)
plot(mlr_admission)
```

Output

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.24985 -0.02360  0.01162  0.03550  0.16638

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.1638772  0.1498288  -7.768 1.64e-13 ***
GRE.Score          0.0016548  0.0007311   2.263 0.024395 *
TOEFL.Score        0.0019967  0.0013344   1.496 0.135716
University.Rating  0.0041870  0.0056870   0.736 0.462219
SOP               -0.0018987  0.0064401  -0.295 0.768353
LOR                0.0246979  0.0069045   3.577 0.000411 ***
CGPA               0.1201825  0.0143089   8.399 2.52e-15 ***
Research           0.0385064  0.0093925   4.100 5.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06437 on 272 degrees of freedom
Multiple R-squared:  0.7988,    Adjusted R-squared:  0.7936
F-statistic: 154.2 on 7 and 272 DF,  p-value: < 2.2e-16
```

MLR 모델의 adjusted Rsquared value 가 1 에 가까울수록 선형성을 띈다. 이 데이터셋의 adjusted Rsquared value 는 0.7936 으로 선형성을 띄지만, 강하지 않다는것을 알 수 있다.





Residual vs Fitted plot 을 관찰해보면, 잔차의 평균값을 나타내는 붉은 지표가 어느정도 평평하기에 이 데이터셋은 homoskedacity 를 만족한다다.

OLS 방식의 solution 이 만족해야 하는 가정은, Noise 가 정규분포를 따라야 한다는 것이다. Normal Q-Q plot 을 관찰해보면, -1 부터 1 구간까진 noise 가 선형성을 띄지만, 그 이외의 구간들은 띄지 않는다. 결론적으로 선형성을 띄지 않기 때문에, 정규분포를 따르지 않다고 여겨진다.

[Q7]

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.1638772  0.1498288  -7.768 1.64e-13 ***
GRE.Score          0.0016548  0.0007311   2.263 0.024395 *
TOEFL.Score        0.0019967  0.0013344   1.496 0.135716
University.Rating  0.0041870  0.0056870   0.736 0.462219
SOP               -0.0018987  0.0064401  -0.295 0.768353
LOR                0.0246979  0.0069045   3.577 0.000411 ***
CGPA               0.1201825  0.0143089   8.399 2.52e-15 ***
Research           0.0385064  0.0093925   4.100 5.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06437 on 272 degrees of freedom
Multiple R-squared:  0.7988,    Adjusted R-squared:  0.7936
F-statistic: 154.2 on 7 and 272 DF,  p-value: < 2.2e-16
```

유의수준 0.01 에서 통계적으로 유의미한 변수들은 GRE, LOR, CGPA, 그리고 Research 이다. 해당 변수들은 전부 양의 상관관계를 가진다. 즉, 이 변수들의 관측치가 1 increment 씩 증가한다면, 해당 변수의 estimate 양 만큼 종속변수 또한 증가한다. 특히 CGPA 의 estimate 은 0.12 로, 강한 양의 상관관계에 있다.

[Q8]

R code

```
mlr_admission_haty <- predict(mlr_admission, newdata = admission_val_data)

perf_mat[1,] <- perf_eval_reg(admission_val_data$Chance.of.Admit, mlr_admission_haty)
perf_mat
```

Output

```
                          RMSE         MAE      MAPE
Admission Predict 0.06393031 0.04740435 7.445885
```

MAE 값을 통해 평균적으로 석사과정 합격 확률의 오차는 0.047 이며, MAPE 값을 통해 오차율은 약 7.4%인걸 알 수 있다.

[Q9]
7 개의 변수를 3 개로 줄인다면, 다음의 가정으로 변수를 선택 할 수 있다.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.1638772  0.1498288  -7.768 1.64e-13 ***
GRE.Score          0.0016548  0.0007311   2.263 0.024395 *
TOEFL.Score        0.0019967  0.0013344   1.496 0.135716
University.Rating  0.0041870  0.0056870   0.736 0.462219
SOP               -0.0018987  0.0064401  -0.295 0.768353
LOR                0.0246979  0.0069045   3.577 0.000411 ***
CGPA               0.1201825  0.0143089   8.399 2.52e-15 ***
Research           0.0385064  0.0093925   4.100 5.47e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

각 변수들의 Pvalue 를 가정했을때, 유의확률이 0 에서 0.001 사이인 LOR, CGPA, Research 변수를 선택하면 된다. 나아가서 correlation graph 를 살펴보면 이 변수들이 다른 변수들과 독립성을 띄는지 살펴본다. LOR 같은 경우는 chance of admit 보다 SOP 와의 연관성이 짙고 research 는 chance of admit 보다 GRE 와 연관성이 있지만, 아주 크게 차이나는 정도가 아니니 이 세개의 변수를 선택하겠다.

[Q10]

R code

```r
id_idx2 <- c(2,3,4,5)

admission_data2 <- cbind(admission_data[,-c(id_idx2)])

nStud2 <- nrow(admission_data2)
nVar2 <- ncol(admission_data2)
set.seed(12345)
admission_trn_idx2 <- sample(1:nStud2, round(0.7*nStud2))
admission_trn_data2 <- admission_data2[admission_trn_idx2,]
admission_val_data2 <- admission_data2[-admission_trn_idx2,]
mlr_admission2 <- lm(Chance.of.Admit ~ ., data = admission_trn_data2)

mlr_admission2
summary(mlr_admission2)
plot(mlr_admission2)

perf_mat2 <- matrix(0, nrow=1, ncol=3)

rownames(perf_mat2) <- c("3 Variables")
colnames(perf_mat2) <- c("RMSE", "MAE", "MAPE")
perf_mat2

mlr_admission_haty2 <- predict(mlr_admission2, newdata = admission_val_data2)

perf_mat2[1,] <- perf_eval_reg(admission_val_data2$Chance.of.Admit, mlr_admission_haty2)
perf_mat2
```

Output

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.250678 -0.027001  0.009833  0.037527  0.198425

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.783346   0.070534 -11.106  < 2e-16 ***
LOR          0.026185   0.006086   4.302 2.35e-05 ***
CGPA         0.161329   0.009679  16.667  < 2e-16 ***
Research     0.048452   0.009098   5.326 2.09e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06574 on 276 degrees of freedom
Multiple R-squared:  0.787,     Adjusted R-squared:  0.7847
F-statistic:   340 on 3 and 276 DF,  p-value: < 2.2e-16
```

```
                    RMSE        MAE      MAPE
3 Variables 0.06806317 0.05173635 8.065257
                    RMSE        MAE      MAPE
Admission Predict 0.06393031 0.04740435 7.445885
```

변수를 LOR, CPGA, Research 3 개로 제한했을때, adjusted Rsquared 값은 0.7847 로 모든 변수를 사용했을때 나오는 adjusted Rsquared 값 0.7936 보다 약 0.0089 작다. 오히려 상대적으로 ‘무의미’ 한 변수들을 지우니 선형성이 낮아진다는것을 볼 수 있다.
또한, 설명변수를 3 개로 제한하니 MAE 값이 0.0043 만큼 증가했고, MAPE 값이 0.61%증가한것을 볼 수 있다. 만약 모든 변수를 사용하는 것과 선택된 변수를 사용하는 것 중 골라야 한다면, adjusted Rsquared 값이 더 높고 MAE MAPE 값이 더 낮은 전자를 사용해야 한다.

[Extra Question]
추가적인 분석으로, 다중공선성 분석과 ANOVA 분석을 실행하였다. 다중공선성은 회귀분석에서 사용된 모형의 일부 설명변수가 다른 설명변수와 상관정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상이다. 다중공선성 분석을 통해 어떤 변수가 다른 변수와 상관계수가 높은지 판단하여, 그 변수를 지움으로써 더 좋은 분석을 할 수 있다.
ANOVA 분석은 각변수의 회귀계수 t 검정을 통해 유의미한 변수인지 아닌지 판단할 수 있게 해준다.

R code

```r
#Extra Question
install.packages("car")
library(car)
vif(mlr_admission)
anova(mlr_admission)
```

다중공선성 분석

```
> vif(mlr_admission)
       GRE.Score        TOEFL.Score University.Rating               SOP               LOR              CGPA          Research
        4.608673          4.391847          2.927981          2.982894          2.528668          4.761596          1.475518
```

R 의 vif 함수를 이용하면 다중공선성 값을 알 수 있는데, 이때 값이 4 를 넘으면 그 변수는 다중공선성을 띈다고 볼 수 있다. 변수들 중 GRE, TOEFL, CGPA 가 다중공선성을 띈다고 할 수 있다.

ANOVA 분석

```
Analysis of Variance Table

Response: Chance.of.Admit
                   Df Sum Sq Mean Sq F value    Pr(>F)
GRE.Score           1 3.5353  3.5353 853.246 < 2.2e-16 ***
TOEFL.Score         1 0.2023  0.2023  48.814 2.166e-11 ***
University.Rating   1 0.1867  0.1867  45.050 1.114e-10 ***
SOP                 1 0.0475  0.0475  11.452 0.0008191 ***
LOR                 1 0.1353  0.1353  32.655 2.886e-08 ***
CGPA                1 0.2969  0.2969  71.658 1.611e-15 ***
Research            1 0.0696  0.0696  16.808 5.467e-05 ***
Residuals         272 1.1270  0.0041
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA 분석을 통하여 각 변수의 유의확률(P value)을 검정할 수 있다. 모든 변수의 유의확률은 0 과 0.001 사이이기 때문에, 모든 변수는 유의미하다고 할 수 있다.