1. Data

Python sklearn dataset에 있는 보스턴의 집값 데이터를 활용하였다. 해당 데이터는 1978년 보스턴 교외의 506개의 주택에 대해 14개 범주의 수치를 정리한 것이다. 14개의 범주는 다음과 같다.

CRIM	도시의 1인당 범죄율
ZN	25,000 평방 피트가 넘는 주택 비율
INDUS	타운당 비소매 사업 면적 비율
CHAS	길이 찰스강과 경계하면 1, 아니면 0
NOX	일산화질소 농도 (0.1ppm 단위)
RM	주택 당 평균 방 수
AGE	1940년 이전에 건축된 주인이 거주하는 주택 비율
DIS	보스턴의 5개 고용센터까지 가중치 거리
RAD	방사형 고속도로로의 접근성 지수
TAX	$10,000 당 최대 재산세율
PTRATIO	타운별 학생-교사 비율
B	아프리카계 미국인의 비율
LSTAT	저소득층 계층의 비율
MEDV	**주인이 거주하는 주택 가격의 중간값 ($1,000 단위)**

(각 변수의 설명, 노랑색 배경이 종속 변수)

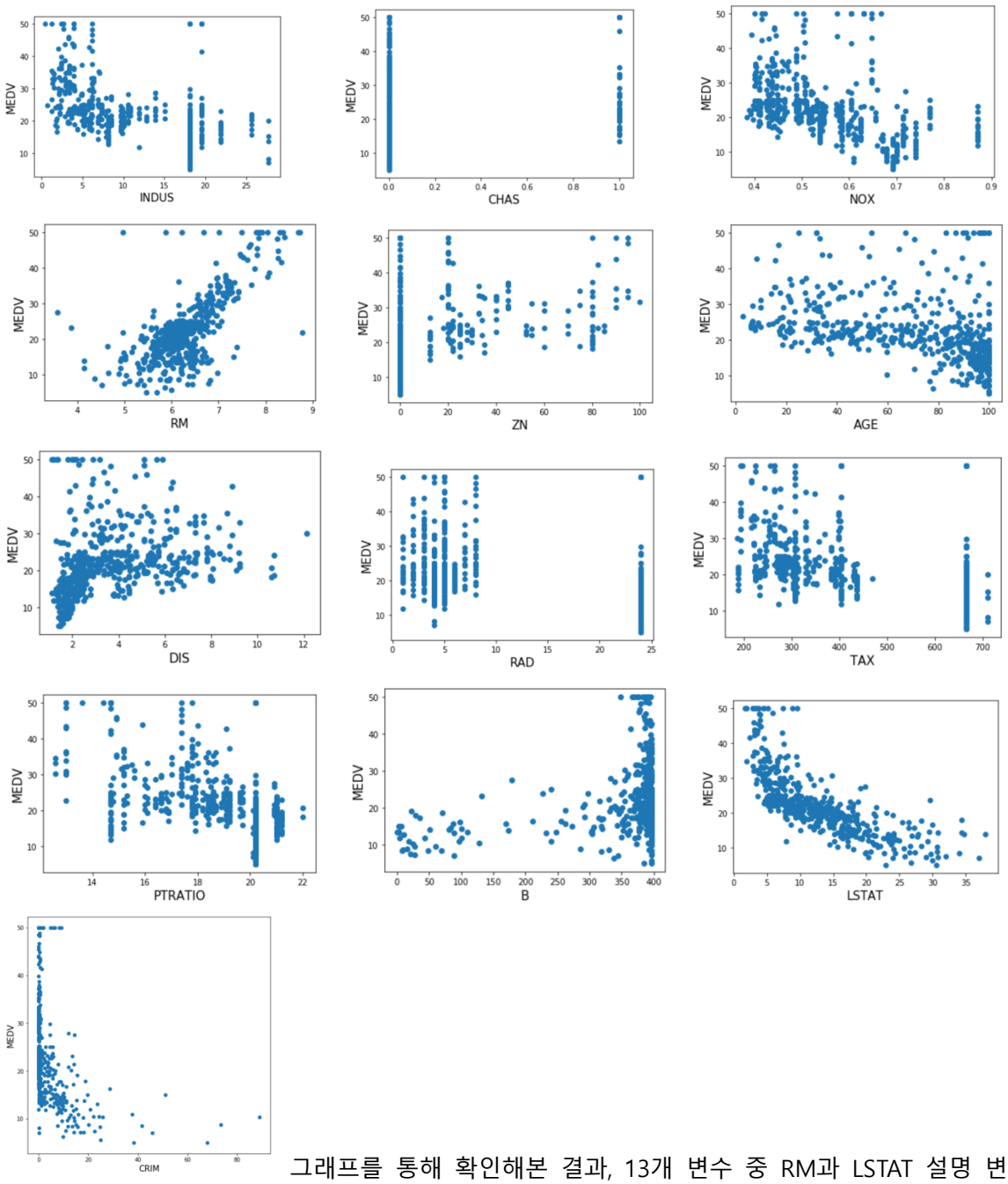
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

해당 data는 총 506개의 데이터로 구성되어 있다. 총 14개의 변수 중 설명변수는 13개이며, 이 중 12개는 수치형 변수, 범주형 변수 1개(CHAS)로 구성되어 있다. 먼저, 각 설명변수와 종속변수 간의 상관관계를 분석하였으며, 그 이후 변수선택법을 진행하여 최적의 회귀모형을 찾아보았다. UCI 머신러닝 리포지토리의 와인 품질 데이터를 사용하였다.

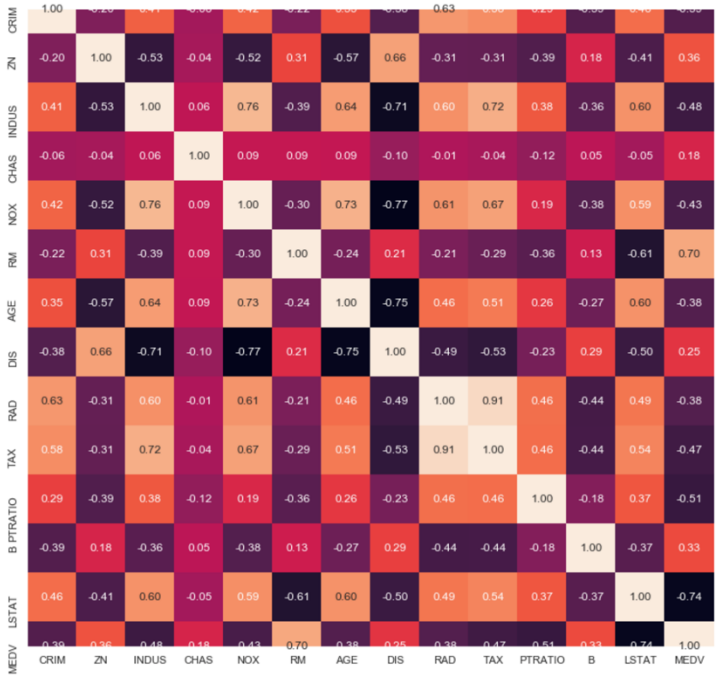
2. 상관관계 분석

본격적인 회귀 모형 추정에 앞서, 각 설명변수와 종속변수간 관계 선형성을 확인하였다. 결과적으로 13개 변수 중 RM과 LSTAT의 선형성이 높은 것으로 파악되었다.

1) **그래프 시각화와 상관계수 계산을 통한 확인**



그래프를 통해 확인해본 결과, 13개 변수 중 RM과 LSTAT 설명 변수 총 2개가 선형 관계를 보였다. 주택 당 방 개수를 의미하는 RM은 종속변수와 정비례 관계로 볼 수 있었고, 저소득층의 비율을 의미하는 LSTAT 변수의 경우 집값이라는 종속변수와 반비례 관계임을 파악할 수 있었다. 더 정밀한 분석을 위하여 상관계수를 계산해보았다.



상관관계의 시각화를 위하여 히트맵을 이용하여 정량적으로 표시해보았다. 양의 상관도가 높으면 색깔이 옅어지고, 음의 상관도가 높으면 색깔이 짙어지고, 아무런 상관관계가 없을 수록 빨강색에 가까운 색으로 표현하였다. 그래프에서 확인할 수 있는 것과 동일하게 상관계수도 RM과 LSTAT 두가지 설명변수에서 각각 RM은 0.70, LSTAT는 -0.74로 상관도가 높게 나왔음을 확인할 수 있었다.

2) **RM, LSTAT의 r^2 계산과 종속변수 간 그래프 도출**

두 가지 상관계수가 높은 변수를 도출하여 OLS를 통해 각 변수의 r^2를 확인하였다.

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.484
Model:	OLS	**Adj. R-squared:**	0.483
Method:	Least Squares	**F-statistic:**	471.8
Date:	Thu, 18 Mar 2021	**Prob (F-statistic):**	2.49e-74
Time:	22:14:03	**Log-Likelihood:**	-1673.1
No. Observations:	506	**AIC:**	3350.
Df Residuals:	504	**BIC:**	3359.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-34.6706	2.650	-13.084	0.000	-39.877	-29.465
RM	9.1021	0.419	21.722	0.000	8.279	9.925

Omnibus:	102.585	**Durbin-Watson:**	0.684
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	612.449
Skew:	0.726	**Prob(JB):**	1.02e-133
Kurtosis:	8.190	**Cond. No.**	58.4

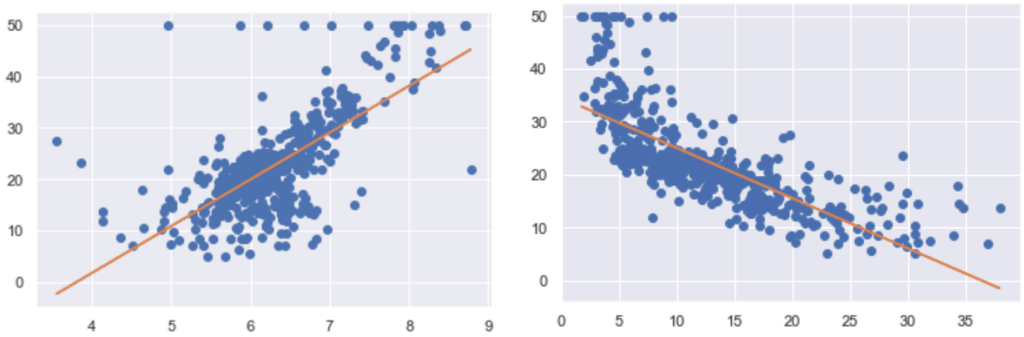
OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.544
Model:	OLS	**Adj. R-squared:**	0.543
Method:	Least Squares	**F-statistic:**	601.6
Date:	Thu, 18 Mar 2021	**Prob (F-statistic):**	5.08e-88
Time:	20:34:11	**Log-Likelihood:**	-1641.5
No. Observations:	506	**AIC:**	3287.
Df Residuals:	504	**BIC:**	3295.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	34.5538	0.563	61.415	0.000	33.448	35.659
LSTAT	-0.9500	0.039	-24.528	0.000	-1.026	-0.874

Omnibus:	137.043	**Durbin-Watson:**	0.892
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	291.373
Skew:	1.453	**Prob(JB):**	5.36e-64
Kurtosis:	5.319	**Cond. No.**	29.7

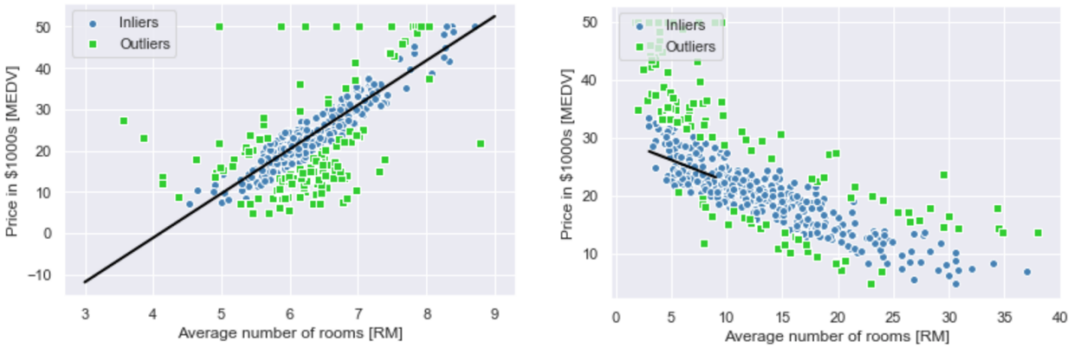
RM 변수의 결정계수는 0.484가 나왔고, LSTAT의 결정계수는 0.544가 나온 것으로 확인하였다. 이후, 두 변수의 각각 종속변수간의 그래프를 그려보았다.



좌측은 RM의 그래프를 나타내고 있으며, 기울기는 약 9.10, 절편은 -34.67로 계산이 되었다. 또한, 우측은 LSTAT의 -0.95의 기울기와 34.55의 절편이 계산되었다.

- **RANSAC을 사용하여 안정된 회귀 모델 훈련**

  → 좀 더 안정적인 각 두 설명변수의 선형회귀모형을 구하기 위하여, outlier를 제거한 샘플들로 회귀 모델을 훈련하는 방식은 RANSAC을 사용하였다. 이 방식은 데이터에서 임의의 개수를 선택하여 이를 inlier로 가정하고 회귀 모델을 구한 후, 구해진 모델에서 다른 모든 데이터들과 비교하고, 지정한 허용 오차 내에 있는 데이터들을 정상치로 포함시킨다. 이후 재구성된 정상치를 이용하여 다시 회귀모델을 구한 이후, 회귀모델과 정산치 간의 오차를 측정한다. 오차가 사용자가 지정한 범위 안에 들어간다면 종료하고 아니라면 반복하는 과정이다.



최대 반복 횟수를 100번으로 제한하였고, 정상치로 포함시키기 위한 허용 오차를 5.0으로 설정하여 해당 모델 훈련을 진행하였다. RANSAC을 통하여 Outlier를 제거하고 도출된 그래프는 다음과 같다. 각각 기울기와 절편 수치가 변화하였다. 방 개수를 의미하는 RM의 경우 기존 기울기 9.10, 절편 -34.67에서 기울기 10.735, 절편 -44.089로 변화하였다. 또한 저소득층 비율을 의미하는 LSTAT의 경우엔 기존 기울기 -0.95, 절편 34.55에서 기울기 -0.746, 절편 29.882로 변화하였다.

3) 상관관계 분석 결과

개별 설명변수의 종속변수 간 상관도 분석 결과, '방 개수'와 정비례의 관계, '저소득층 비율'과는 반비례 관계로 파악이 되었다. 이를 통해 전반적으로 방 개수가 높을 수록 집값은 높게 책정되고, 저소득층 비율이 높을 수록 집값이 감소하는 관계를 볼 수 있었다. 하지만 아웃라

이어의 존재로 반례 또한 발견할 수 있었다.

3. Variable Selection

13개의 설명변수로 집값을 설명하는 모델을 적합해 보았다. 총 6가지의 변수선택 기법인 forward selection, backward elimination, stepwise selection, ridge regression, lasso regression, elastic net regression을 이용해 분석해 보았다. 본격적인 변수 선택에 앞서 전체 13개 설명변수에 대한 다중선형회귀를 진행하였다.

1) 전체 설명 변수에 대한 다중선형회귀

13개의 설명변수에 대한 다중 선형회귀를 진행한 결과, 각 변수의 coefficient는 다음과 같이 구하였다.

```
[-1.04108047e-01  4.00636095e-02  5.62622740e-02  2.79033761e+00
 -1.79507724e+01  4.61934795e+00 -4.98438080e-03 -1.31806109e+00
  2.95903098e-01 -1.28319670e-02 -9.84286378e-01  8.58479134e-03
 -4.33529288e-01]
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.959
Model:                            OLS   Adj. R-squared (uncentered):              0.958
Method:                 Least Squares   F-statistic:                              891.3
Date:                Sat, 20 Mar 2021   Prob (F-statistic):                        0.00
Time:                        14:49:05   Log-Likelihood:                         -1523.8
No. Observations:                 506   AIC:                                      3074.
Df Residuals:                     493   BIC:                                      3128.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0929      0.034     -2.699      0.007      -0.161      -0.025
x2             0.0487      0.014      3.382      0.001       0.020       0.077
x3            -0.0041      0.064     -0.063      0.950      -0.131       0.123
x4             2.8540      0.904      3.157      0.002       1.078       4.630
x5            -2.8684      3.359     -0.854      0.394      -9.468       3.731
x6             5.9281      0.309     19.178      0.000       5.321       6.535
x7            -0.0073      0.014     -0.526      0.599      -0.034       0.020
x8            -0.9685      0.196     -4.951      0.000      -1.353      -0.584
x9             0.1712      0.067      2.564      0.011       0.040       0.302
x10           -0.0094      0.004     -2.395      0.017      -0.017      -0.002
x11           -0.3922      0.110     -3.570      0.000      -0.608      -0.176
x12            0.0149      0.003      5.528      0.000       0.010       0.020
x13           -0.4163      0.051     -8.197      0.000      -0.516      -0.317
==============================================================================
Omnibus:                      204.082   Durbin-Watson:                   0.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1374.225
Skew:                           1.609   Prob(JB):                     3.90e-299
Kurtosis:                      10.404   Cond. No.                      8.50e+03
==============================================================================
```

기본적인 OLS를 이용한 분석 결과 값이다. 다중선형회귀 데이터들을 Train과 Test 데이터로 구분하여, 현재 데이터의 상황을 살펴보였다.

```
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

print('전체 변수 훈련 MSE: %.3f, 전체 변수 테스트 MSE: %.3f' % (
        mean_squared_error(y_train, y_train_pred),
        mean_squared_error(y_test, y_test_pred)))
print('전체 변수 훈련 R^2: %.3f, 전체 변수 테스트 R^2: %.3f' % (
        r2_score(y_train, y_train_pred),
        r2_score(y_test, y_test_pred)))
```

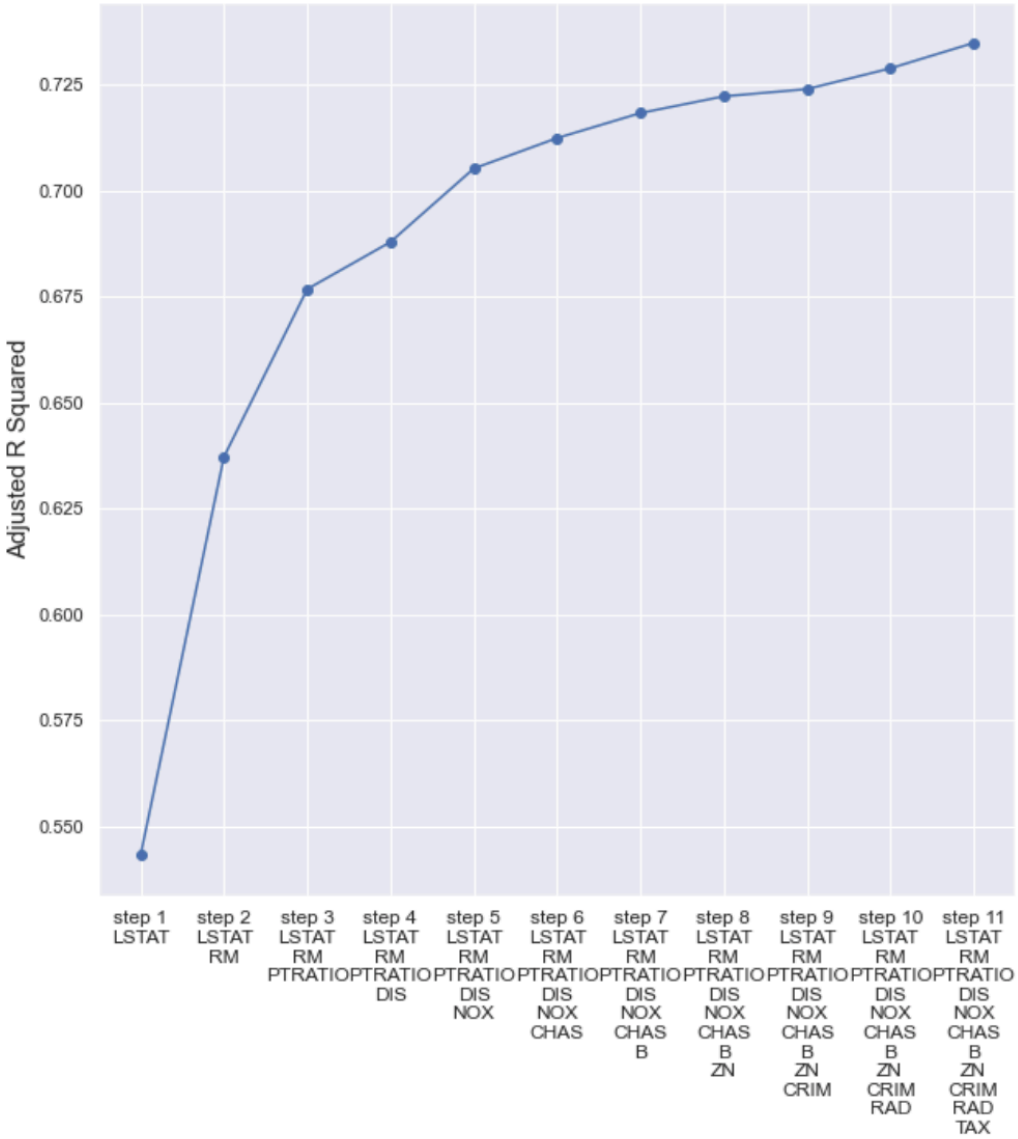
```
전체 변수 훈련 MSE: 21.037, 전체 변수 테스트 MSE: 26.564
전체 변수 훈련 R^2: 0.747, 전체 변수 테스트 R^2: 0.703
```

결과적으로, 전체 변수에 대한 다중회귀를 진행한 결과 테스트 데이터의 MSE가 더 높은 것으로 보아 overfitting 되어 있다는 것으로 파악하였고, Overfitting을 해결하기 위하여 변수선택을 진행하였다.

2) **Variable Selection**

2-1) Forward Selection

유의수준을 0.1로 설정한 후, forward selection을 진행하였다. 분석결과 첫번째 스탭에서 LSTAT이 선택되었고, 두번째에서는 RM이 선택되었다. 이는 위 상관도를 분석한 결과, 두 변수가 가장 상관도가 높기 때문에 상관도의 크기에 따라 선택된 것으로 보인다. 또한 비소매 업종의 비율을 의미하는 INDUS와, 오래된 주택 비율을 의미하는 AGE가 제외된 것으로 확인되었는데, 이 두가지는 보스턴의 집값에 다른 변수 대비 크게 영향을 주지 않는 것으로 판단할 수 있었다.



그래프 분석 결과 단계가 추가될 때마다, 결정계수의 변화량이 줄어든 것으로 확인할 수 있

었다. 특히, 9단계에서 넘어갈 때인 CRIM 변수가 추가될 때는 가장 적게 결정계수가 변화하는 것으로 볼 수 있었다.

OLS Regression Results

Dep. Variable:	MEDV	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.958
Method:	Least Squares	F-statistic:	1057.
Date:	Sat, 20 Mar 2021	Prob (F-statistic):	0.00
Time:	14:49:06	Log-Likelihood:	-1523.9
No. Observations:	506	AIC:	3070.
Df Residuals:	495	BIC:	3116.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.4254	0.048	-8.917	0.000	-0.519	-0.332
x2	5.8979	0.296	19.909	0.000	5.316	6.480
x3	-0.3951	0.109	-3.631	0.000	-0.609	-0.181
x4	-0.9298	0.177	-5.262	0.000	-1.277	-0.583
x5	-3.3945	3.096	-1.097	0.273	-9.477	2.688
x6	2.8248	0.897	3.151	0.002	1.063	4.586
x7	0.0149	0.003	5.531	0.000	0.010	0.020
x8	0.0498	0.014	3.508	0.000	0.022	0.078
x9	-0.0928	0.034	-2.702	0.007	-0.160	-0.025
x10	0.1743	0.064	2.724	0.007	0.049	0.300
x11	-0.0096	0.004	-2.710	0.007	-0.016	-0.003

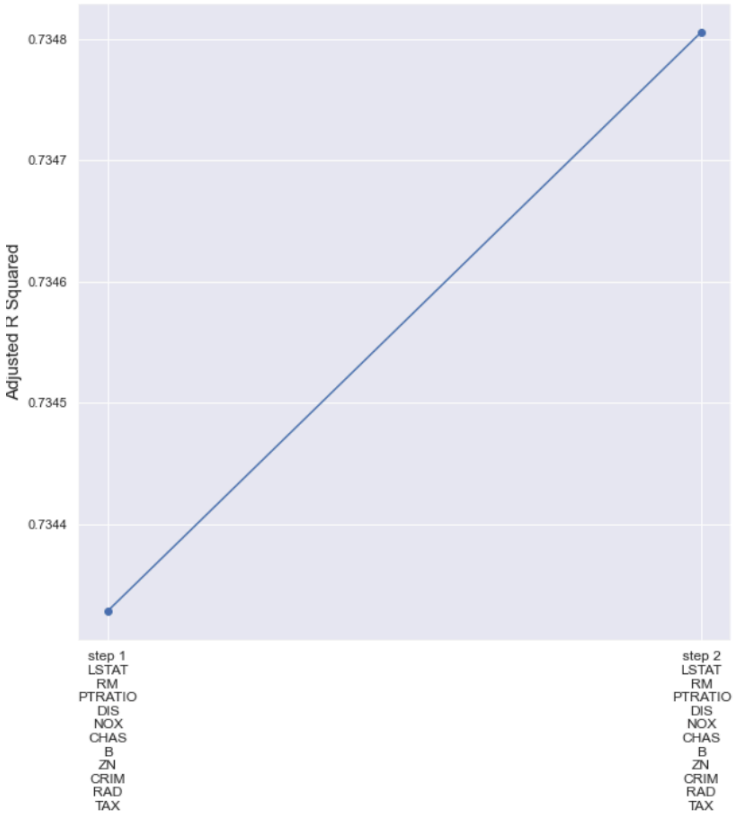
Omnibus:	200.446	Durbin-Watson:	1.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1313.904
Skew:	1.584	Prob(JB):	4.89e-286
Kurtosis:	10.230	Cond. No.	7.79e+03

다음과 같은 각 변수의 계수를 구할 수 있었다. 이에 따른 회귀식은 다음과 같다.

$$MEDV = -0.4254LSTAT + 5.8979RM - 0.3951PTRATIO - 0.9298DIS - 3.3945NOX + 2.8248CHAS + 0.0149B + 0.0498ZN - 0.0928CRIM + 0.1743RAD - 0.0096TAX$$

2-2) Backward Selection

유의수준을 0.1로 설정한 후, backward elimination을 진행하였다. 분석 결과, 선택된 변수는 앞선 Forward Selection과 동일한 변수들이 선택된 것으로 도출되었다.



OLS Regression Results

Dep. Variable:	MEDV	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.958
Method:	Least Squares	F-statistic:	1057.
Date:	Sat, 20 Mar 2021	Prob (F-statistic):	0.00
Time:	14:49:06	Log-Likelihood:	-1523.9
No. Observations:	506	AIC:	3070.
Df Residuals:	495	BIC:	3116.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.0928	0.034	-2.702	0.007	-0.160	-0.025
x2	0.0498	0.014	3.508	0.000	0.022	0.078
x3	2.8248	0.897	3.151	0.002	1.063	4.586
x4	-3.3945	3.096	-1.097	0.273	-9.477	2.688
x5	5.8979	0.296	19.909	0.000	5.316	6.480
x6	-0.9298	0.177	-5.262	0.000	-1.277	-0.583
x7	0.1743	0.064	2.724	0.007	0.049	0.300
x8	-0.0096	0.004	-2.710	0.007	-0.016	-0.003
x9	-0.3951	0.109	-3.631	0.000	-0.609	-0.181
x10	0.0149	0.003	5.531	0.000	0.010	0.020
x11	-0.4254	0.048	-8.917	0.000	-0.519	-0.332

Omnibus:	200.446	Durbin-Watson:	1.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1313.904
Skew:	1.584	Prob(JB):	4.89e-286
Kurtosis:	10.230	Cond. No.	7.79e+03

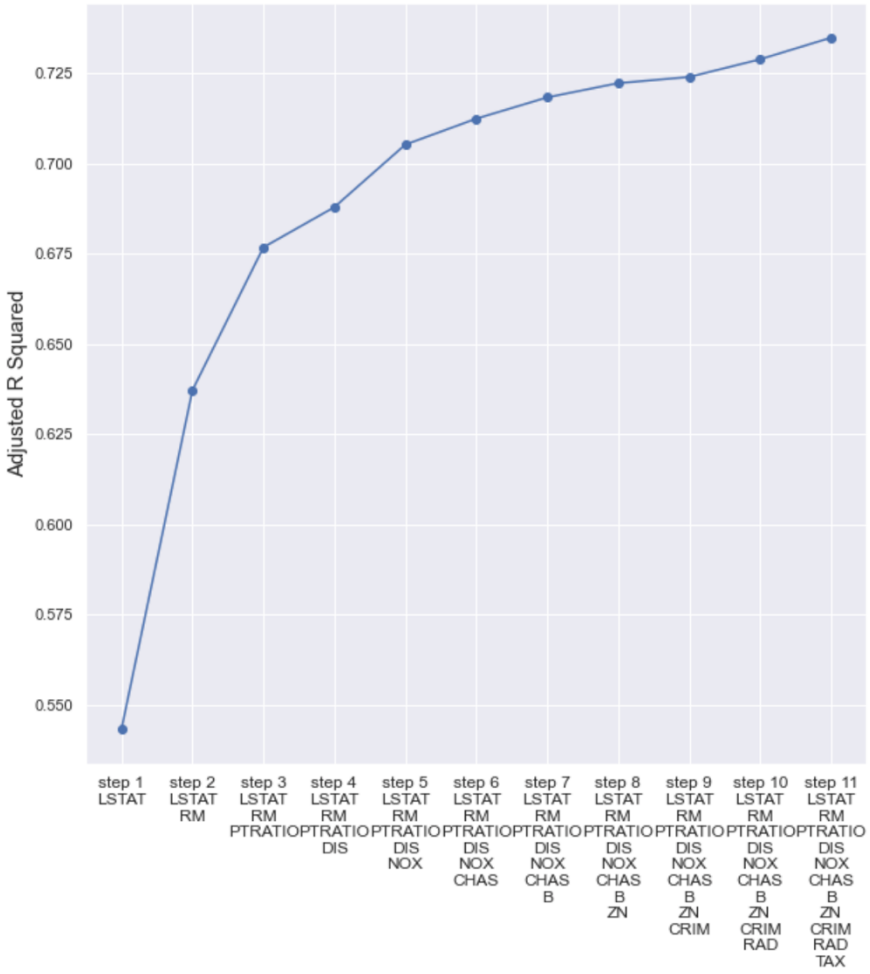
앞선 방식과 같은 변수가 선택되었기 때문에, 결정계수 값은 변화하지 않았고, 회귀식의 계수 또한 변화하지 않았다. 따라서 Backward Selection 방식을 통한 도출된 회귀식은 다음과 같다.

$$MEDV = -0.4254LSTAT + 5.8979RM - 0.3951PTRATIO - 0.9298DIS - 3.3945NOX + 2.8248CHAS + 0.0149B + 0.0498ZN - 0.0928CRIM + 0.1743RAD - 0.0096TAX$$

2-3) Stepwise Selection

유의수준을 0.1로 설정한 후, Stepwise Seletion을 진행하였다. 분석 결과, 선택된 변수는 앞

선 Forward Selection, Backward Elemination과 동일한 변수들이 선택된 것으로 도출되었다.



단계마다 소거 과정 없이 추가 되는 과정을 보여 그래프는 Forward Selection과 동일한 형태를 보였다. Forward Selection에서 마찬가지로 앞선 상관도가 높은 변수인 LSTAT와 RM이 첫 두 단계에서 선택 되었으며, 단계가 거듭될수록 결정계수의 변화량이 작아졌다.

OLS Regression Results

Dep. Variable:	MEDV	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.958
Method:	Least Squares	F-statistic:	1057.
Date:	Sat, 20 Mar 2021	Prob (F-statistic):	0.00
Time:	14:49:07	Log-Likelihood:	-1523.9
No. Observations:	506	AIC:	3070.
Df Residuals:	495	BIC:	3116.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.4254	0.048	-8.917	0.000	-0.519	-0.332
x2	5.8979	0.296	19.909	0.000	5.316	6.480
x3	-0.3951	0.109	-3.631	0.000	-0.609	-0.181
x4	-0.9298	0.177	-5.262	0.000	-1.277	-0.583
x5	-3.3945	3.096	-1.097	0.273	-9.477	2.688
x6	2.8248	0.897	3.151	0.002	1.063	4.586
x7	0.0149	0.003	5.531	0.000	0.010	0.020
x8	0.0498	0.014	3.508	0.000	0.022	0.078
x9	-0.0928	0.034	-2.702	0.007	-0.160	-0.025
x10	0.1743	0.064	2.724	0.007	0.049	0.300
x11	-0.0096	0.004	-2.710	0.007	-0.016	-0.003

Omnibus:	200.446	Durbin-Watson:	1.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1313.904
Skew:	1.584	Prob(JB):	4.89e-286
Kurtosis:	10.230	Cond. No.	7.79e+03

위 방법 역시 동일한 11개의 변수가 선택되었기 때문에, 결정계수와 회귀식의 계수가 다 동일한 값으로 도출되었다. 따라서 Stepwise Selection 방식을 통한 도출된 회귀식은 다음과 같다.

$$MEDV = -0.4254LSTAT + 5.8979RM - 0.3951PTRATIO - 0.9298DIS - 3.3945NOX + 2.8248CHAS + 0.0149B + 0.0498ZN - 0.0928CRIM + 0.1743RAD - 0.0096TAX$$

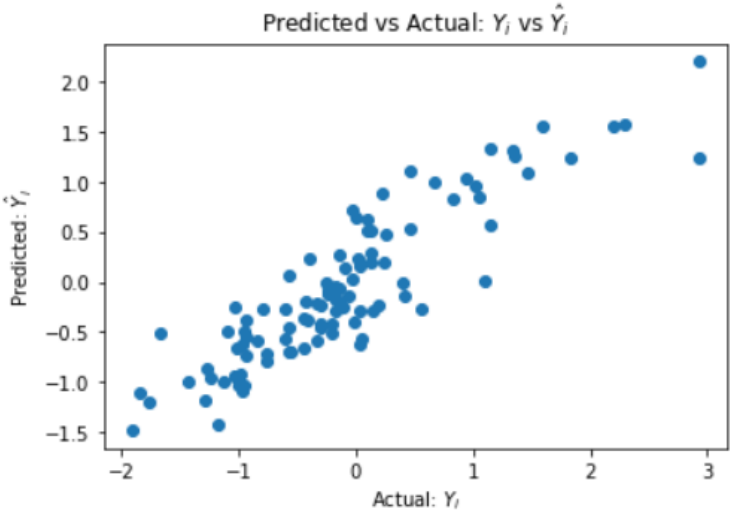
2-4) Ridge Regression

Ridge regression model은 기존선형회귀에서 발생할 수 있는 독립변수들간의 다중공선성 문제의 해결방안으로 제시되었다. 이는 다중선형회귀 모형의 선형 계수값의 크기를 감소시켜 다중공선성에 의한 예측 오차를 줄이게된다. Ridge Regression의 식은 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{N}(y_i - (X\beta)_i)^2 + \lambda\sum_{j=1}^{P}\beta_j^2\right), \quad \text{for } \lambda \geq 0.$$

기존의 β추정식에 penalty항이 추가된것으로, λ는 ridge function의 alpha값이다. Alpha값이 커질수록 penalty가 커지면서 계수의 크기가 줄어들게된다.

ridge regression을 진행하여 그래프상에 plotting을 해보면 다음과 같은 결과가 나온다.

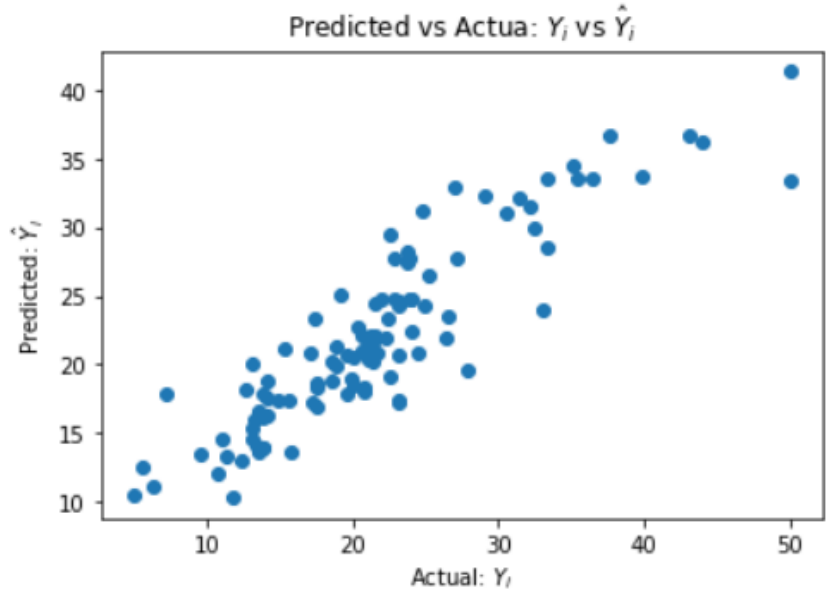


2-5) Lasso Model

Lasso Regression은 ridge회귀의 선형 계수가 0이 될수없어 변수선택이 불가능하다는 단점을 보완하기위해 제안되었다. 이는 계수의 절댓값에 대한 패널티를 부여함으로써 이루어지는 축소추정법이다. Lasso Regression의 식은 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{N}(y_i - (X\beta)_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j| \right), \quad \text{for } \lambda \geq 0.$$

lasso regression을 진행하여 그래프상에 plotting을 해보면 다음과 같은 결과가 나온다.

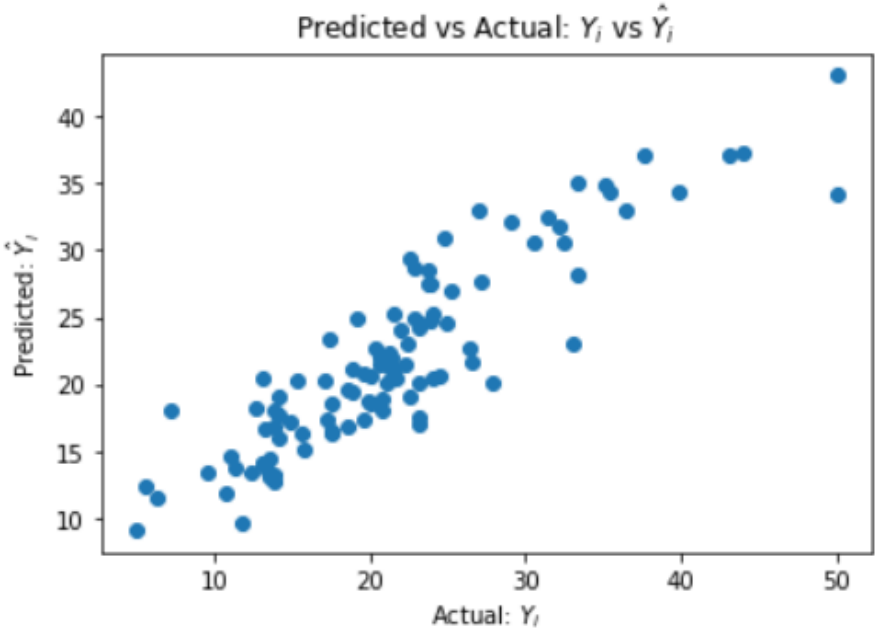


2-6) Elastic net

elastic net은 lasso와 ridge의 절충안으로 나온 모델로 상관관계가 있는 변수들 중에서 하나의 변수만을 흔히 선택하는 lasso의 단점을 보완하는 형태이다. λ_1은 ratio로서 0에 가까울수록 ridge모델에 가까워지며, 1에 가까울수록 lasso모델에 가까워진다. λ_2는 alpha 값이다.

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{N}(y_i - (X\beta)_i)^2 + \lambda_1\sum_{j=1}^{P}|\beta_j| + \lambda_2\sum_{j=1}^{P}\beta_j^2\right), \quad \text{for } \lambda_1, \lambda_2 \geq 0.$$

elastic net regression을 진행하여 그래프상에 plotting을 해보면 다음과 같은 결과가 나온다.



비교

Forward selection, backward elimination, stepwise selection을 했을 때 모두 동일한 11개의 변수가 선택되었다. 세 방식 모두 같은 변수가 선택되어 모두 R^2와 MSE의 값이 동일하게 나왔다.

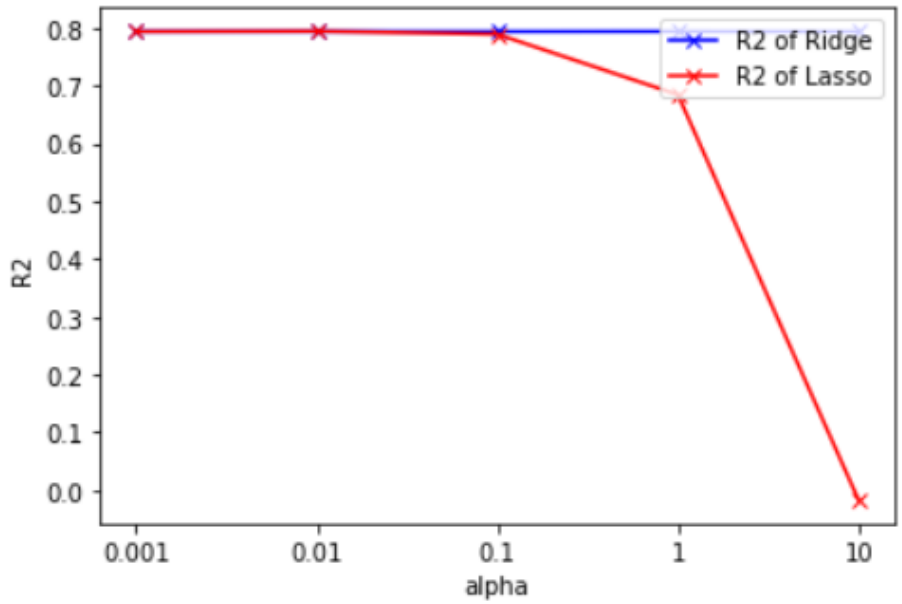
	Forward Selection	**Backward Selection**	**Stepwise Selection**
CRIM	O	O	O
ZN	O	O	O
INDUS	X	X	X
CHAS	O	O	O
NOX	O	O	O
RM	O	O	O
AGE	X	X	X
DIS	O	O	O
RAD	O	O	O
TAX	O	O	O
PTRATIO	O	O	O
B	O	O	O
LSTAT	O	O	O

변수 citric_acid가 제거 되었고 같은 회귀 결과가 나왔다. Ridge를 이용하여 회귀 분석을 한 결과

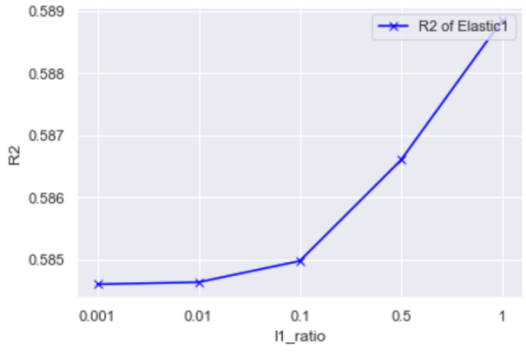
제거되는 변수들은 없었지만, 계수들의 절댓값의 크기가 줄어들었다. Lasso, elastic net을 사용해서 회귀 분석을 한 결과 제거되는 변수가 없었고 OLS와 같은 결과를 보였다. 변수 제거를 통해서 다중공선성을 줄이는 효과보다 있는 설명 변수를 모두 이용하여 설명력을 높이는 효과가 크기 때문인 것 같다. 변수수가 더 많고 변수 간의 다중공선성이 더 컸더라면 제거되는 변수가 생겼을 것이고 ridge, lasso, elastic net을 통해서 더 좋은 설명력을 가진 모델을 적합할 수 있었을 것이라고 생각한다.'

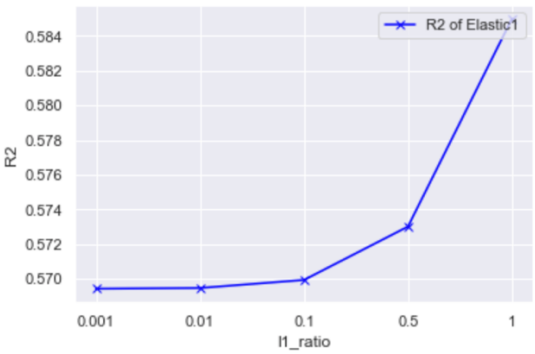
	전체 변수	Forward	Backward	Stepwise	Ridge	Lasso	Elastic
Test R^2	0.703	0.590	0.590	0.590	0.58	0.56	0.58
Train R^2	0.747	0.773	0.773	0.773	0.77	0.76	0.77
Test MSE	26.564	33.424	33.424	33.424	34.23	36.07	33.79
Train MSE	21.037	19.329	19.329	19.329	19.49	20.24	19.36

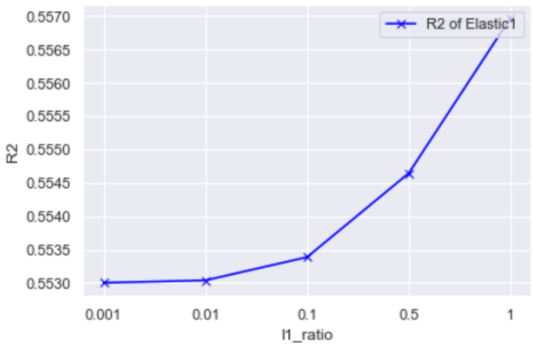
따라서, 나머지 세개의 축소추정 방식을 진행한 결과, ridge회귀와 lasso회귀의 alpha값을 변화시켜보며 r_square값을 구해보면 다음과 같다.

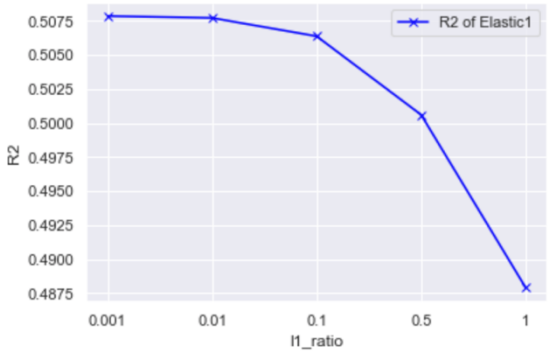


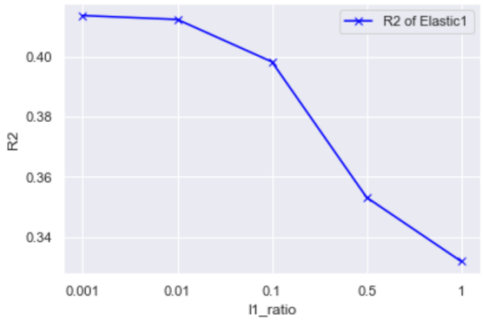
alpha값의 변화함에 따라 Lasso의 R2값이 영향을 크게 받으므로 Lasso모델보다 Ridge모델이 Robust 하다고 할 수 있다. 또 각 alpha값에 따라, 각 ratio값에 따라 변화하는 elastic net모델을 비교해본다면 다음과 같다.











alpha가 0.001에서 10까지 변화하는 상황에서 ratio를 변화시켜보았다. 이때 ratio가 0.5에서 1사이일 때, 즉 lasso에 가까운 모델일 때, R2값이 Robust하지못하다고 판단할 수 있다. 여기에 더해 3가지 모델의 MSE와 R2값을 비교해보면 0과 0.5사이의 ratio를 가진 elastic net을 사용하는게 적합한 모델이라고 생각된다.

4. Chi-square Test

다음으로는 여러가지 독립변수와 종속변수의 관계가 독립적인지를 검정하기 위해 **Chi-square Test**를 진행하였다. 우선, 검정할 독립변수로 '도시에서 소매 업종이 아닌 비율', '주택의 평균 방 개수', '1940년 이전에 지어진 자가 주택 비율', '10만 달러당 재산세율', '다섯 개의 보스턴 고용 센터까지 가중치가 적용된 거리'를 선택해 이 5가지 변수에 대한 이산화를 진행하였다. 5가지 변수 모두 연속형 변수였기 때문에 각각의 최솟값, 최댓값, 평균과 중앙값을 참고해 threshold를 설정한 후 이산화를 진행하였고, 그 결과는 아래의 표와 같다.

변수	값
도시에서 소매 업종이 아닌 비율 (INDUS)	~9.69 / 9.69~
주택의 평균 방 개수 (RM)	~6.2085/6.2085~
1940년 이전에 지어진 자가 주택 비율 (AGE)	~77.5/77.5~
10만 달러당 재산세율 (TAX)	~330/330~
다섯 개의 보스턴 고용 센터까지 가중치가 적용된 거리 (DIS)	~3.20745/3.20745

위 다섯 가지 변수의 모든 경우에 대해 독립변수와의 카이제곱 검정을 진행해본 결과, 각 변수의 p-value 모두 0.05보다 작은 수가 나왔다. 모든 경우에 대해 범주형 독립변수와 종속변수 간의 관계가 없다는 귀무가설 하에서 계산된 검정통계량 값의 p-value가 0.05이하라는 것은, 유의수준 0.05하에서 독립변수와 종속변수가 독립적이라는 귀무가설을 기각하고 두 변수간의 관계가 종속적이라는 대립가설을 채택할 수 있다.

	INDUS	**RM**	**AGE**	**TAX**	**DIS**
검정통계량	77.92	135.21	80.03	96.95	33.37
자유도	1	1	1	1	1
P-Value	<0.05	<0.05	<0.05	<0.05	<0.05