**HW3**

**20200639 Chae Woojin**

**Chapter 6:** **4, 5, 7, 9(excluding (e), (f)), 10**

**Chapter 7:** **1, 2, 4, 10, 11**

**Chapter 6**

4.

(a) iii. As increases, model flexibility decreases. This leads to the increase of training RSS.

(b) ii. Test RSS will decrease until decreasing variance overwhelms increasing bias. Then, it will bounce back as effect of increasing bias wins the effect of decreasing variance.

(c) iv. As increases, the model becomes less flexible, so the variance steadily decreases.

(d) iii. However, in terms of bias, it steadily increases as model becomes less flexible,

(e) v. The irreducible error is unchanged.

5.

(a)

(b)

Let . Then,

.

A minimum can be found when these are set to 0.

As , solution of and is same.

(c)

(d)

Let . Then, as we can simplify partial derivative in regard to each coefficient as

,

where , denotes subdifferential of and respectively.

Then, as partial derivatives contain the subgradient in its term, it is well known that

where

is evaluated similarly.

From this point, one can conclude that we cannot denote its solution uniquely.

7.

(a)

(b)

The posterior probability can be calculated by multiplying the prior and likelihood.

(c)

Mode of can be achieved by finding which maximizes the likelihood of posterior probability.

.

As first term is independent of , our solution will be when we maximize the second term.

As and if we set , the mode corresponds to lasso optimization.

(d)

(e)

To show that the ridge estimate is the mode, we can again find the maximum by maximizing the log of the posterior probability. The log is

As the first term is independent of , it suffices to consider finding that minimizes .

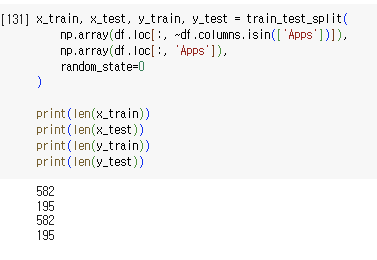
As if we set , this term becomes , which is ridge estimation. In sum, mode corresponds to the ridge regression estimate for under this posterior distribution.

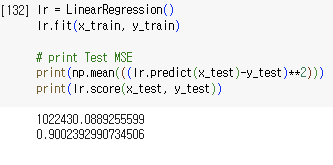
Next, to show that ridge estimate is the mean of the posterior probability, lets note that posterior probability is also gaussian function, because both likelihood and prior are also gaussian.

It is well known fact that mode=maximum=mean in gaussian. Therefore, mean also corresponds to the ridge regression estimate for under this posterior distribution.

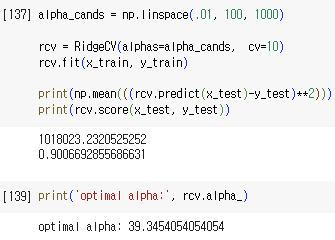
9.

(a)



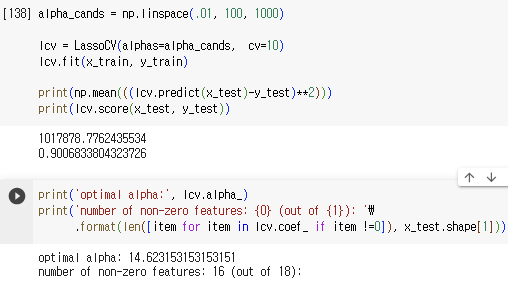
(b)

Test error was about 1022430.

(c)

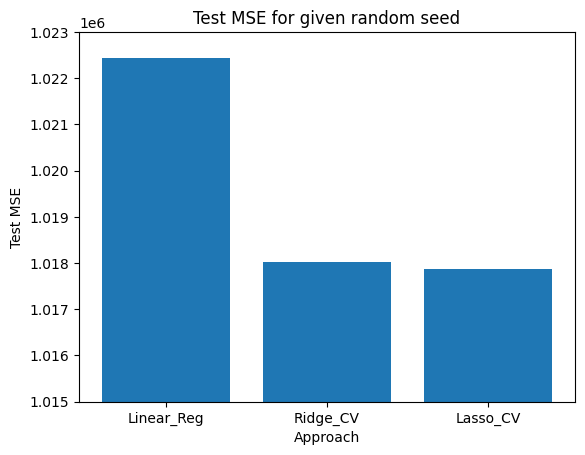
Optimal lambda chosen by cross validation was 39.345, and test error was about 1018023.

(d)



Optimal lambda chosen by cross validation was 14.623, and test error was about 1017878. Also, the number of non-zero coefficient estimates were 16 out of 18.

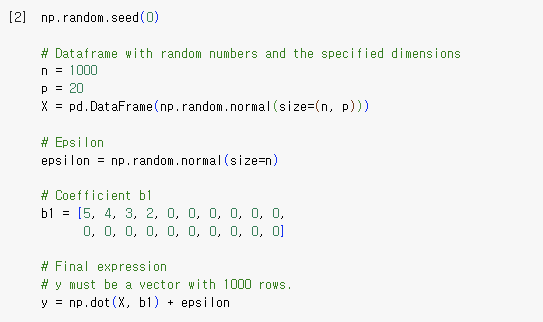
(g)



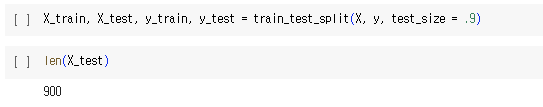
Smallest test error was obtained by lasso model in this given random seed. Ridge model also gave comparable performance against normal linear regression. I think this is because these shrinkage models were known to reduce variance of the estimates by leaving explanatory terms (removing colinear estimates or noise terms), so that they are likely to work better on the new data.

10.

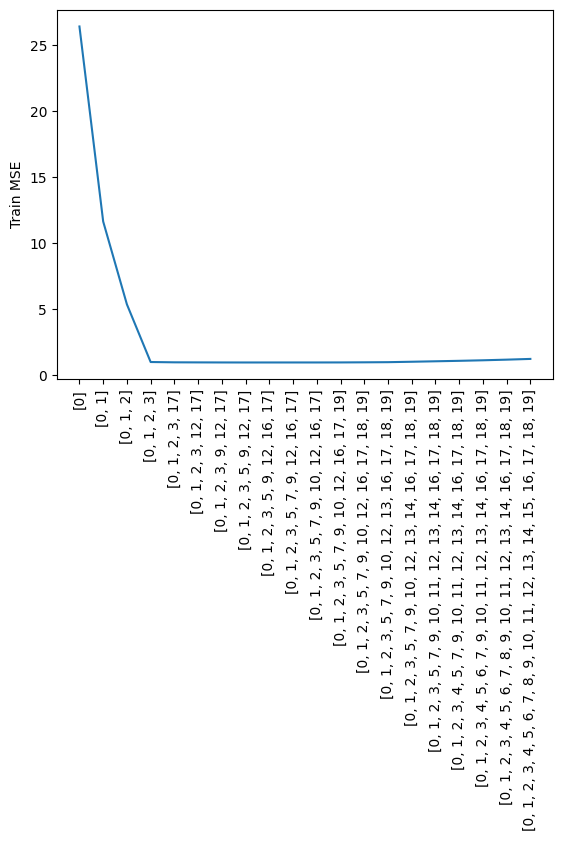
(a)



(b)

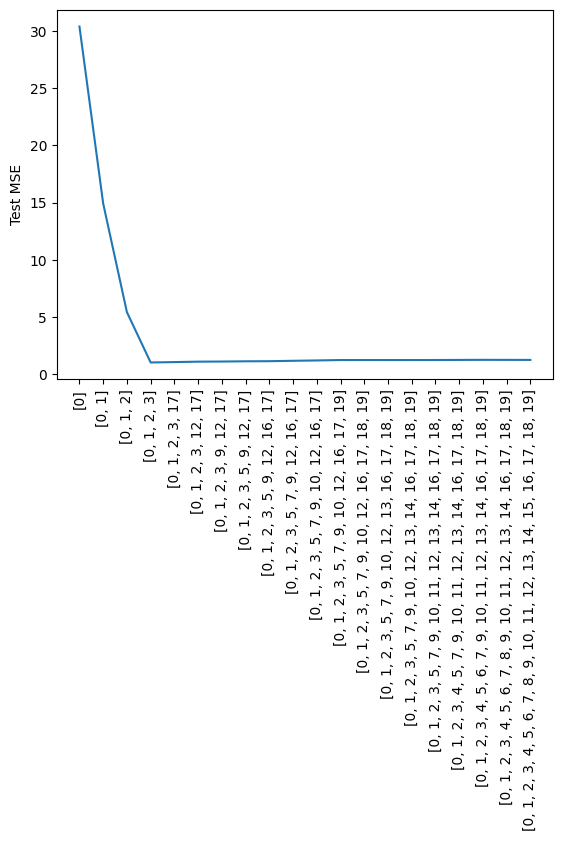


(c)



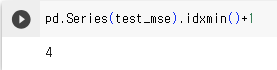
X-tick represents the selected best model.

(d)



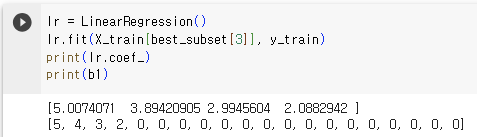
X-tick represents the selected best model.

(e)

****

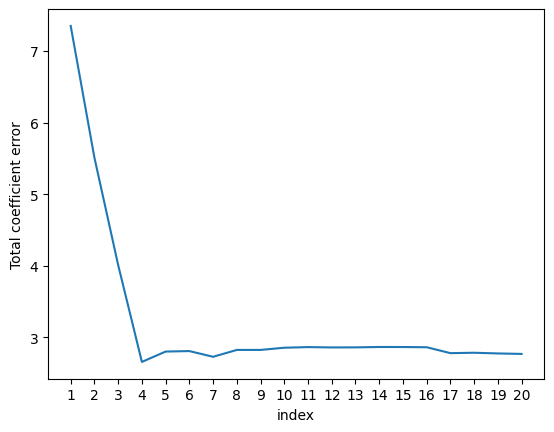
The minimum test MSE is found when model size is 4. This corresponds to the true data which has 4 non-zero values. (I added one in the result because index starts with 0.)

(f)



Coefficient values between two model (best model’s coefficients vs real coefficients) resembles clearly.

(g)



The total error of coefficient estimates is minimized when model size is 4. It corresponds to the size where test MSE is minimized.

**Chapter 7**

1.

(a)

with

For , set , , , , then

(b)

For , set , , , , then

(c)

,

, is continuous at .

(d)

,

,

, is continuous at .

(e)

,

,

, is continuous at .

2.

(a)

As penalty term of original function dominates, should be zero.

(b)

Still penalty term of first derivative dominates, so that slope of should be zero. Hence, is a constant function that minimizes (i.e,)

(c)

Still penalty term of second derivative dominates, so that slope of should be zero. Hence is a linear function that minimizes

(d)

Similarly, penalty term of third derivative of function dominates, so that slop of should be zero. Hence, we can estimate that will be a quadratic function or a linear function.

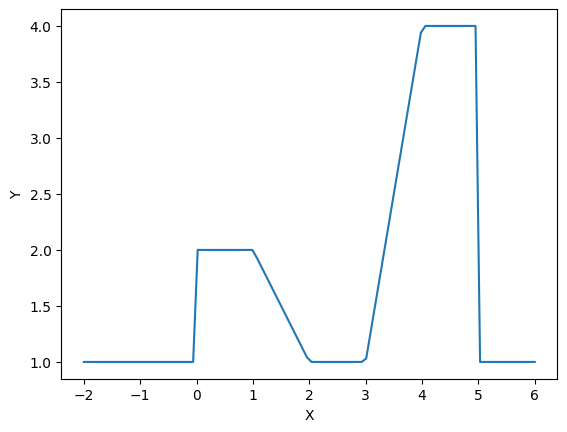
(e)

As we do not penalize third derivative of function, it is same with finding optimal function that minimizes RSS. Therefore, will be able to interpolate all points.

4.

, (Assume for simplicity)

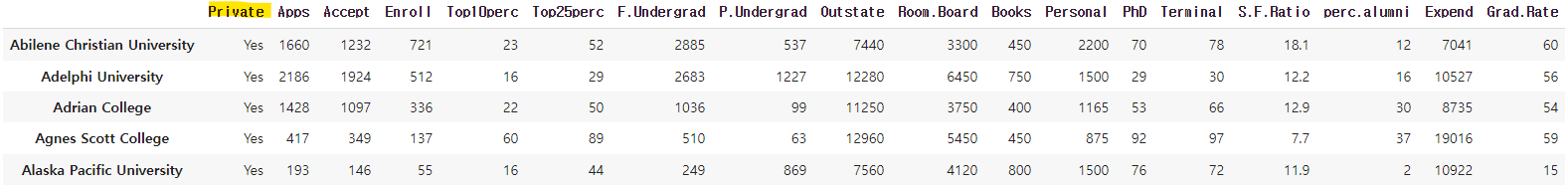
Then, with

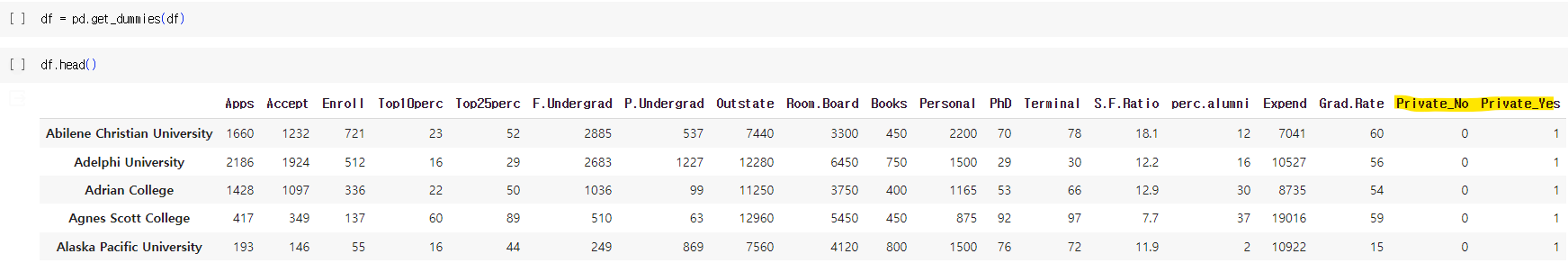


10.

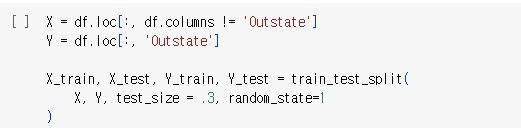
(a)

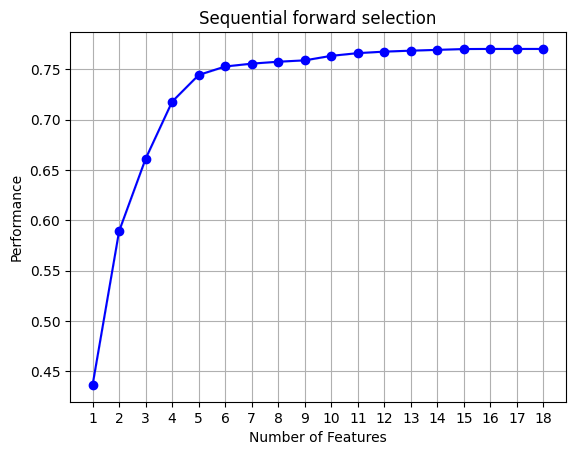
Before splitting given data into train data and test data set, I converted every qualitative predictor into dummy variable (In this data, ‘Private’ was that case.)





Then, I split given data set into train, and test data set.





The picture above describes the R-square value as number of feature increases using forward selection. It seems that taking subset consists up to 6th features (['Room.Board', 'perc.alumni', 'Expend', 'Private\_No', 'PhD', 'Grad.Rate']) work well. However, lets look further using some criteria (Adjusted R-square, Mallow’s Cp, AIC) below.

According to ‘Adjusted R-square’ criteria, ['Room.Board', 'perc.alumni', 'Expend', 'Private\_No', 'PhD', 'Grad.Rate', 'Personal', 'Top25perc', 'Accept', 'Apps', 'Enroll', 'Top10perc', 'S.F.Ratio', 'Terminal', 'Books'] was the best subset of predictors for estimating out-of-state tuition.

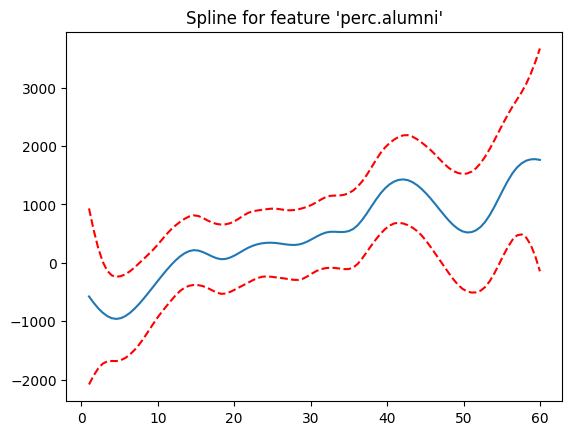
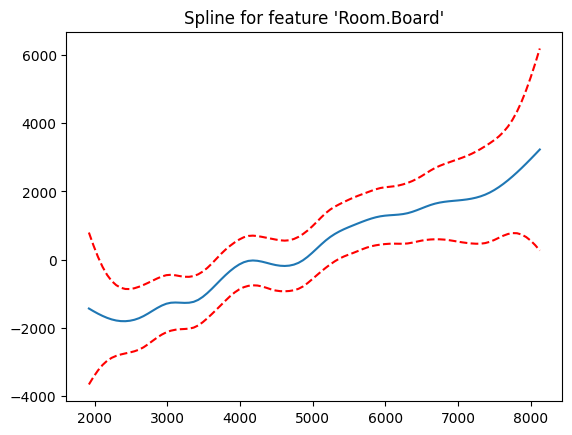
According to ‘Mallow’s Cp’ criteria, ['Room.Board'] was the best subset of predictors for estimating out-of-state tuition.

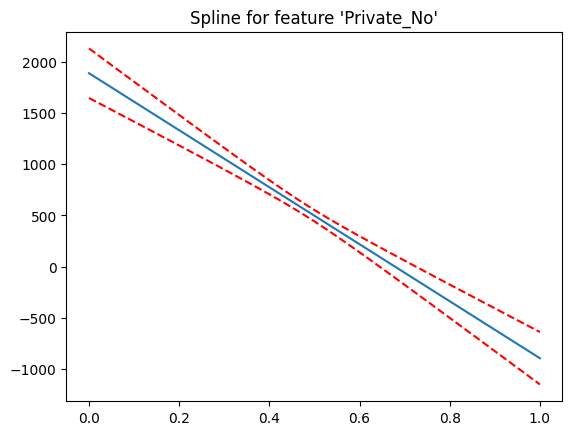
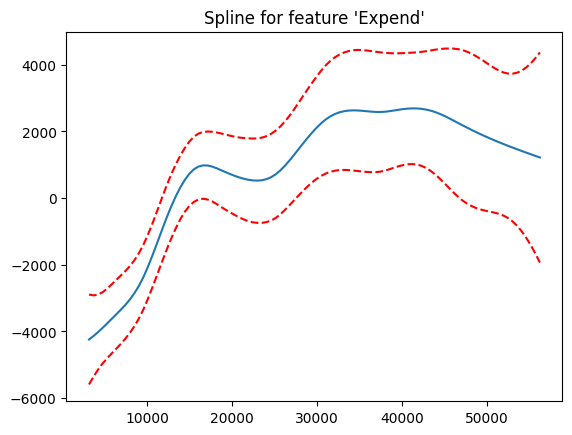
According to ‘AIC’ criteria, ['Room.Board', 'perc.alumni', 'Expend', 'Private\_No', 'PhD', 'Grad.Rate', 'Personal', 'Top25perc', 'Accept', 'Apps', 'Enroll', 'Top10perc', 'S.F.Ratio'] was the best subset of predictors for estimating out-of-state tuition.

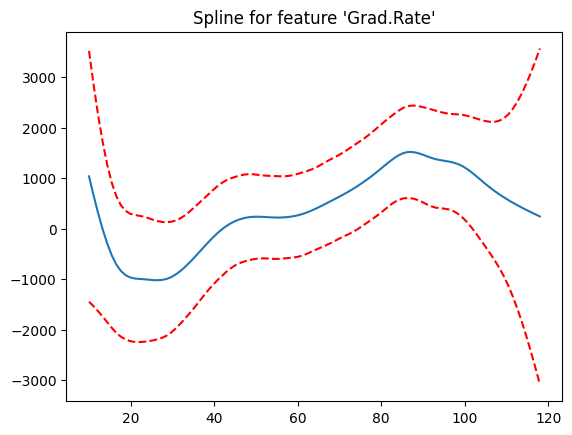
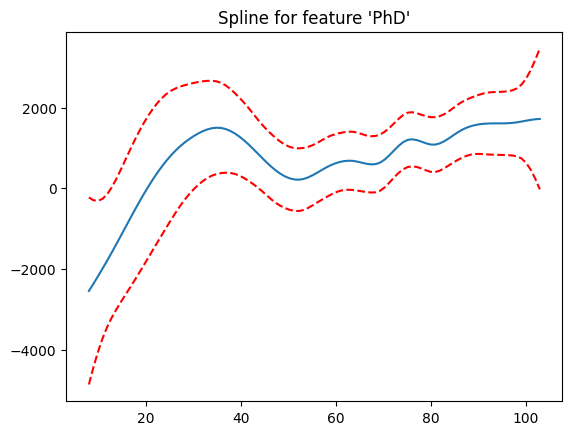
One can easily see that different model was selected based on each criterion.

For the sake of simplicity, I will use top 6 features as a training data set (['Room.Board', 'perc.alumni', 'Expend', 'Private\_No', 'PhD', 'Grad.Rate']), according in the further problem.

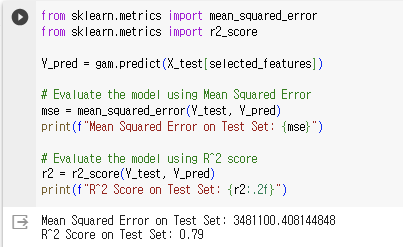
(b)







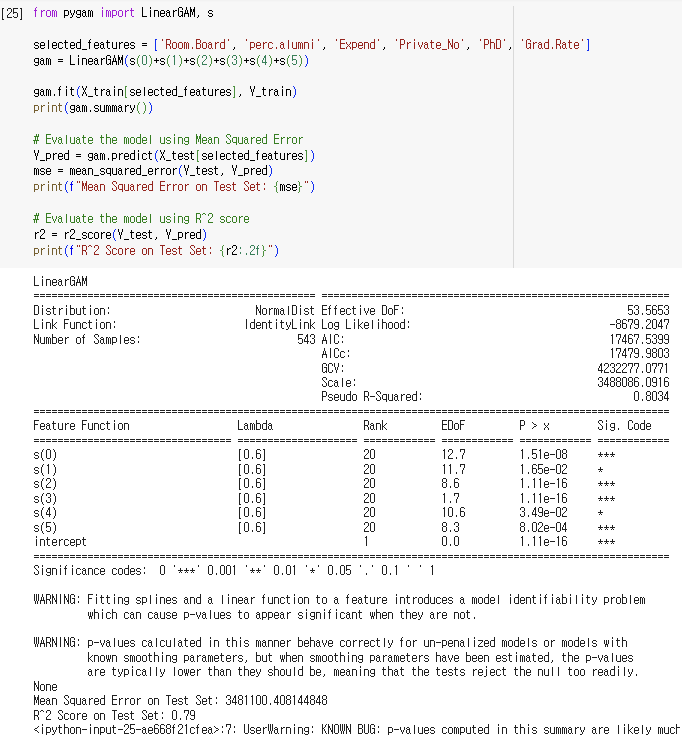
(c)



Test MSE was about 3481100, and R-square value was 0.79.

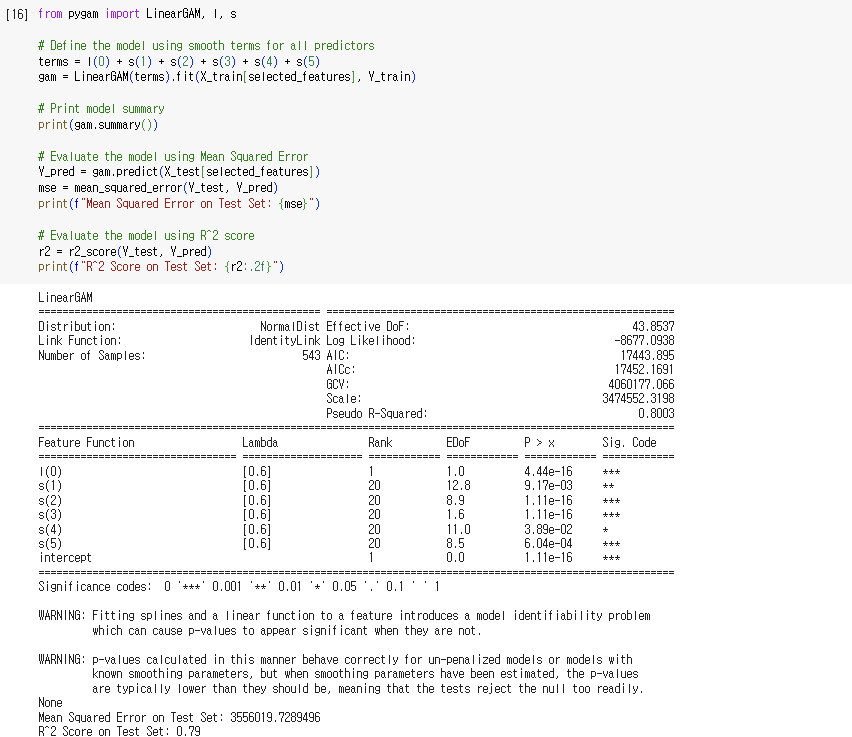
(d)

This picture summarizes the statistics of GAM when every predictor is fitted as spline (non-parametrically).

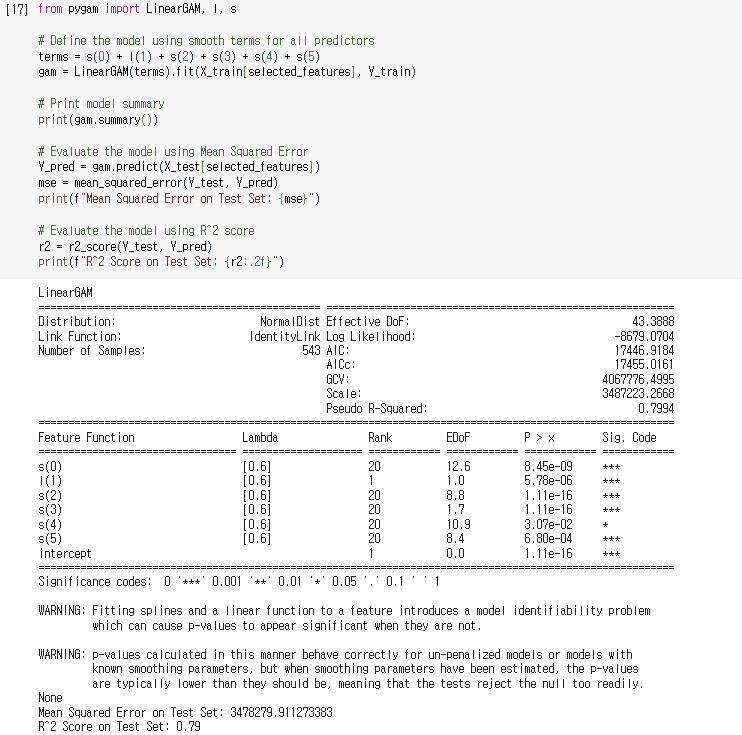


And the six figures below show the results when we substitute each predictor linearly.

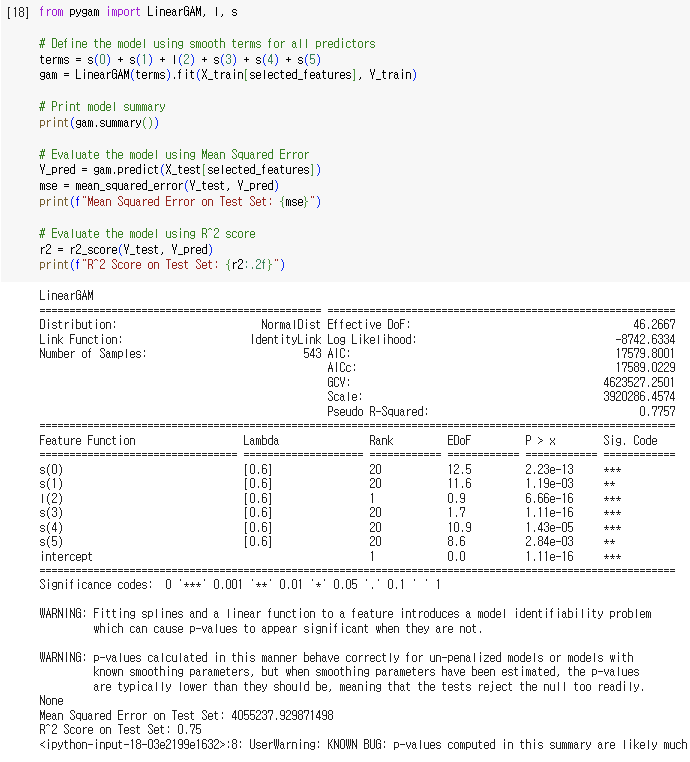
1. Linear term on Room.Board



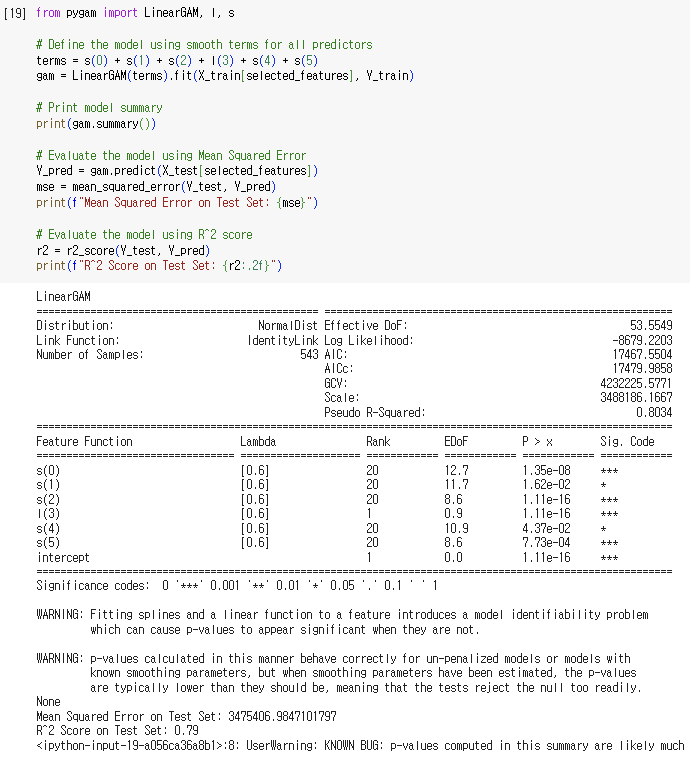
1. Linear term on perc.alumni



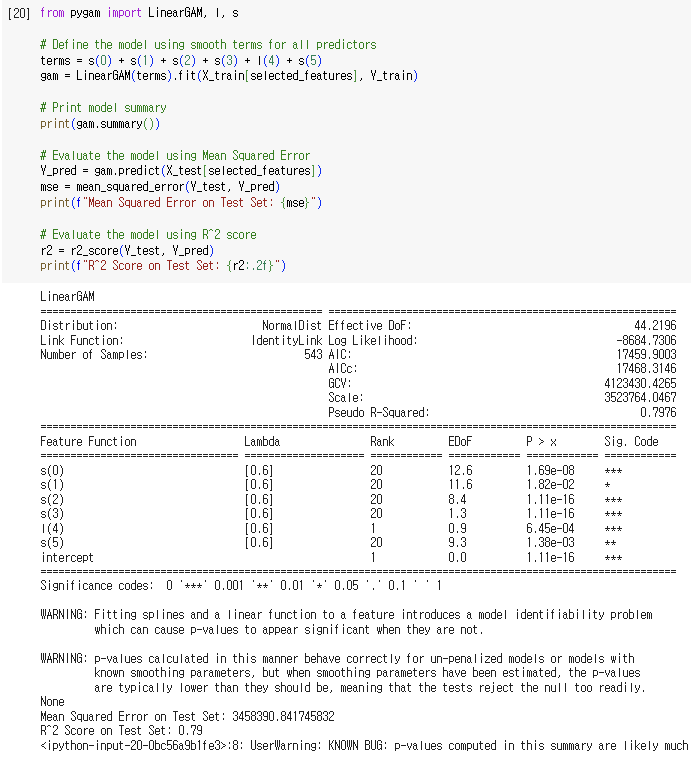
1. Linear term on Expend



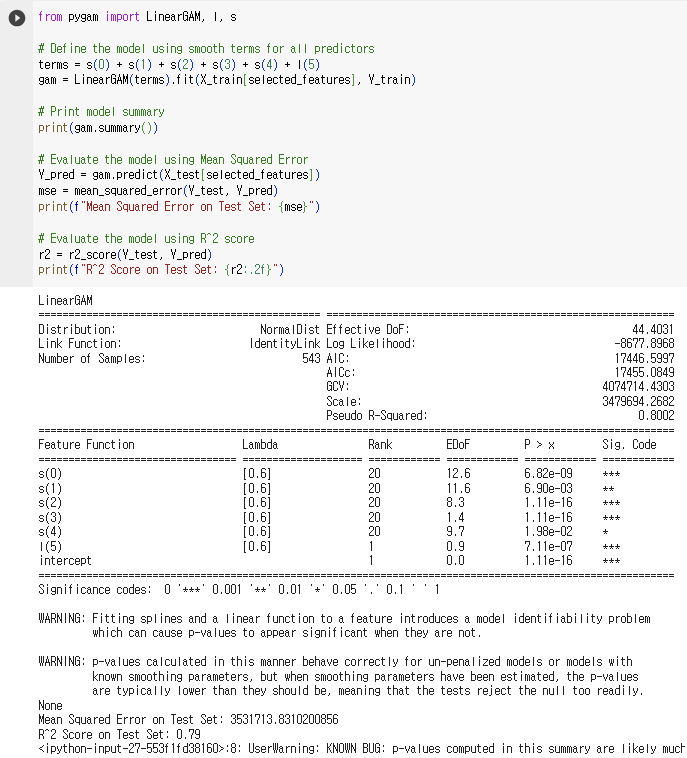
1. Linear term on Private\_No



1. Linear term on PhD



1. Linear term on Grad\_Rate

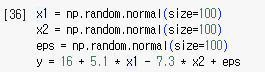


|  |  |  |  |
| --- | --- | --- | --- |
| Model | GCV | Test MSE | AIC |
| s(0)+ s(1)+ s(2)+ s(3)+ s(4)+ s(5) | 4232277 | 3481100 | 17467 |
| l(0)+ s(1)+ s(2)+ s(3)+ s(4)+ s(5) | 4060177 | 3556019 | 17443 |
| s(0)+ l(1)+ s(2)+ s(3)+ s(4)+ s(5) | 4067776 | 3478279 | 17446 |
| s(0)+ s(1)+ **l(2)**+ s(3)+ s(4)+ s(5) | 4623527 | 4055237 | 17579 |
| s(0)+ s(1)+ s(2)+ l(3)+ s(4)+ s(5) | 4232225 | 3475406 | 17467 |
| s(0)+ s(1)+ s(2)+ s(3)+ l(4)+ s(5) | 4123430 | 3458390 | 17459 |
| s(0)+ s(1)+ s(2)+ s(3)+ s(4)+ l(5) | 4074714 | 3531713 | 17446 |

From the statistics, one can conclude that 3rd feature ‘Expend’ has the most significant non-linear relationship with the response, because the model performance worsens the most when we tried to substitute its non-parametric approximation into parametric approximation (linear regression in this case).

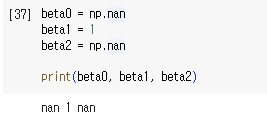
11.

(a)

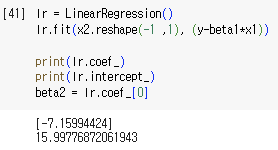


,

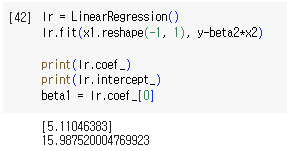
(b)



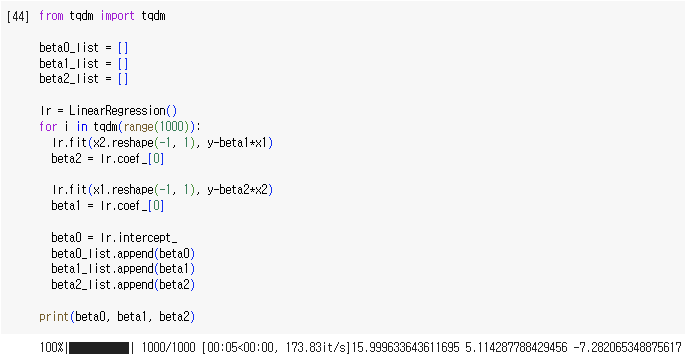
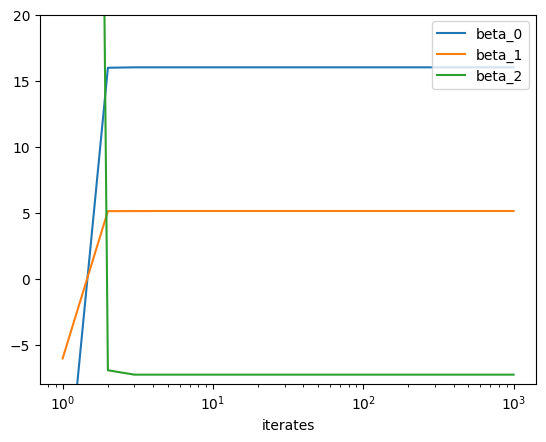
(c)



(d)

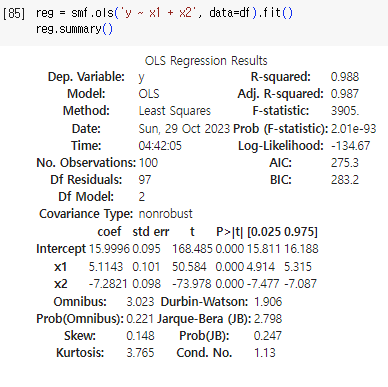
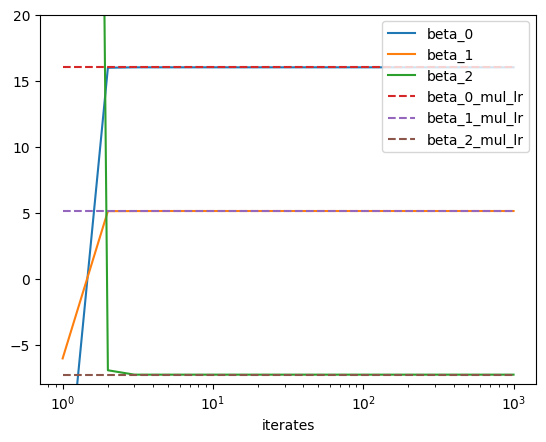


(e)

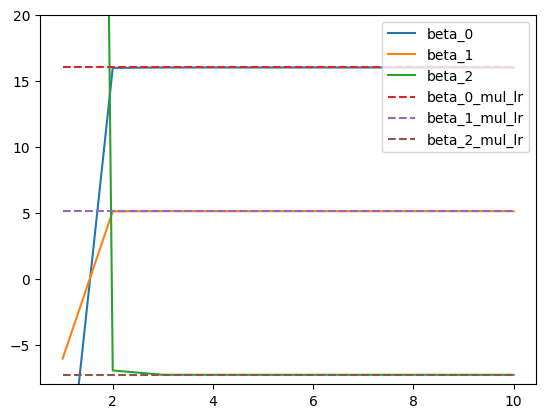
After 1000 iterates, we obtained

(f)

From multiple linear regression, we obtained,

(g)



From the graph visualizing performance of backfitting algorithm up to 10th iterates, it seems that 3~4 iterations are enough to get a good coefficient estimate.