Problem Set 7

MACS 30100 - Perspectives on Computational Modeling Luxi Han 10449918

Problem 1

1.

The following is the regression result:

Table 1: Simple Linear Regression of Biden Warmth

	Dependent variable:
	biden
female	4.103^{***}
	(0.948)
age	0.048^{*}
	(0.028)
dem	15.424^{***}
	(1.068)
rep	-15.850^{***}
	(1.311)
educ	-0.345^{*}
	(0.195)
Constant	58.811^{***}
	(3.124)
Observations	1,807
R^2	0.282
Adjusted R^2	0.280
Residual Std. Error	19.914 (df = 1801)
F Statistic	141.150^{***} (df = 5; 1801)
Note:	*p<0.1; **p<0.05; ***p<0.01

[1] "1 a) The MSE for OLS is: 395.270169278648"

```
##      (Intercept)  female        age      dem       rep       educ
## [1,]    58.81126 4.10323 0.04825892 15.42426 -15.84951 -0.3453348
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  58.81125899  3.1244366  18.822996 2.694143e-72
## female        4.10323009  0.9482286   4.327258 1.592601e-05
## age           0.04825892  0.0282474   1.708438 8.772744e-02
## dem          15.42425563  1.0680327  14.441745 8.144928e-45
## rep         -15.84950614  1.3113624 -12.086290 2.157309e-32
## educ         -0.34533479  0.1947796  -1.772952 7.640571e-02
```

The MSE for the simple linear regression model is 395.27.

Table 2: Simple Linear Regression using Cross Validation of Biden Warmth

	Dependent variable:
	biden
female	4.103***
	(0.948)
age	0.048*
	(0.028)
dem	15.424***
	(1.068)
rep	−15.850***
	(1.311)
educ	−0.345*
	(0.195)
Constant	58.811***
	(3.124)
Observations	1,807
R^2	0.282
Adjusted R^2	0.280
Residual Std. Error	19.914 (df = 1801)
F Statistic	141.150*** (df = 5; 1801)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.

[1] "1 b) The MSE for the cross validation method is: 393.825307793911"

The MSE for the cross validation method is approximately 393.83. The MSE for the cross validation method is slightly lower than the simple linear regression model. But the difference is small, we can't conclude which model is better.

3

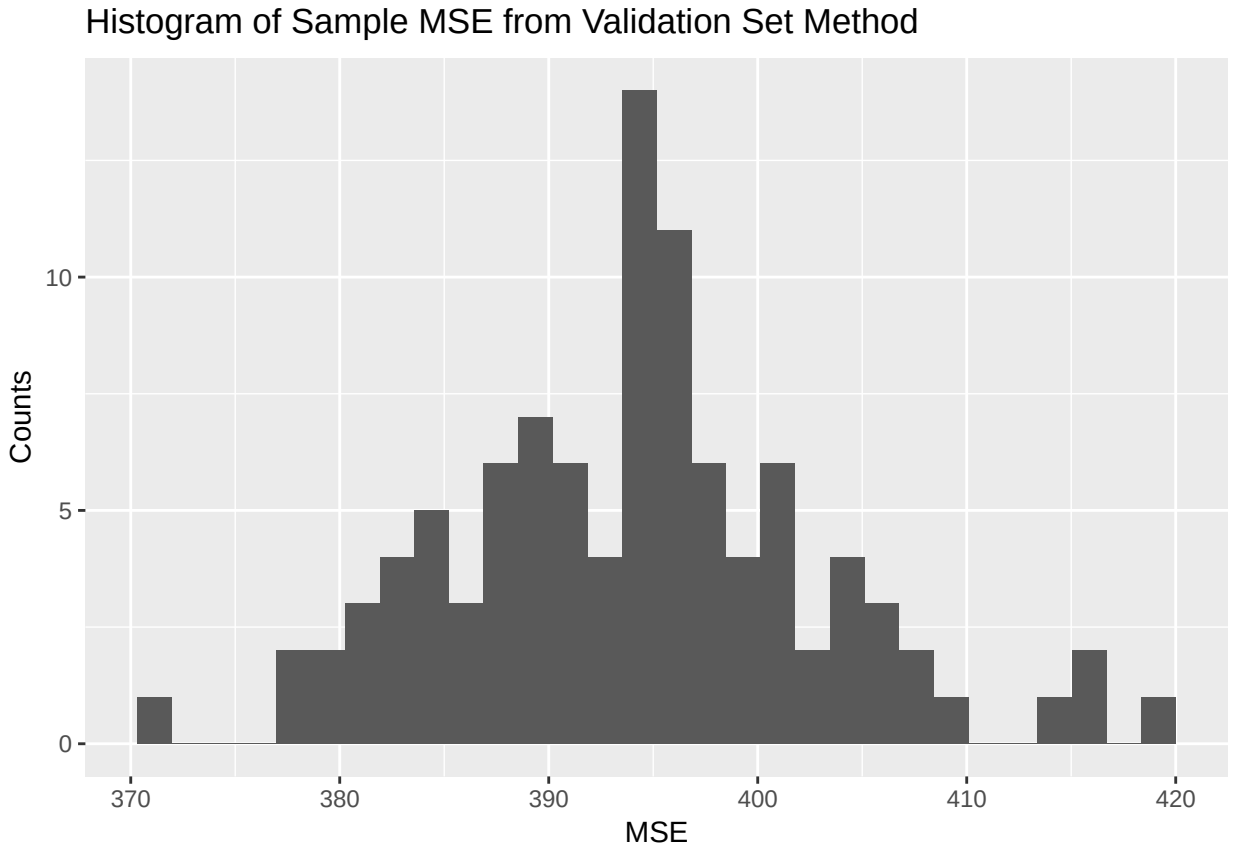
% latex table generated in R 3.3.2 by xtable 1.8-2 package % Mon Feb 27 00:08:30 2017

	colMeans.coef_table.
X.Intercept.	58.81
female	4.10
age	0.05
dem	15.42
rep	-15.85
educ	-0.35
mse	394.05

[1] "Standard deviation of MSE is: 8.82993945501422"

The above is the mean of the 100 validation set estimation results for the coefficients. Additionally, we have the average MSE for the 100 split. The MSE is around 394.05. The estimation is about the same as the regression using the full sample.

We can also plot the histogram of MSE:



4.

```
## [1] 397.9555
```

The leave one out estimation for the MSE is 397.96. This is approximately the same as the full sample approach. This indicates the good fitness of the model.

5

```
## [1] 397.8837
```

The ten fold estimation for the MSE is 397.88. This is almost the same as LOOCV estimation. This corroborates the fact that the performance of ten fold estimation is almost as good as LOOCV, with the advantage that ten fold estimation requires less computation power.

6

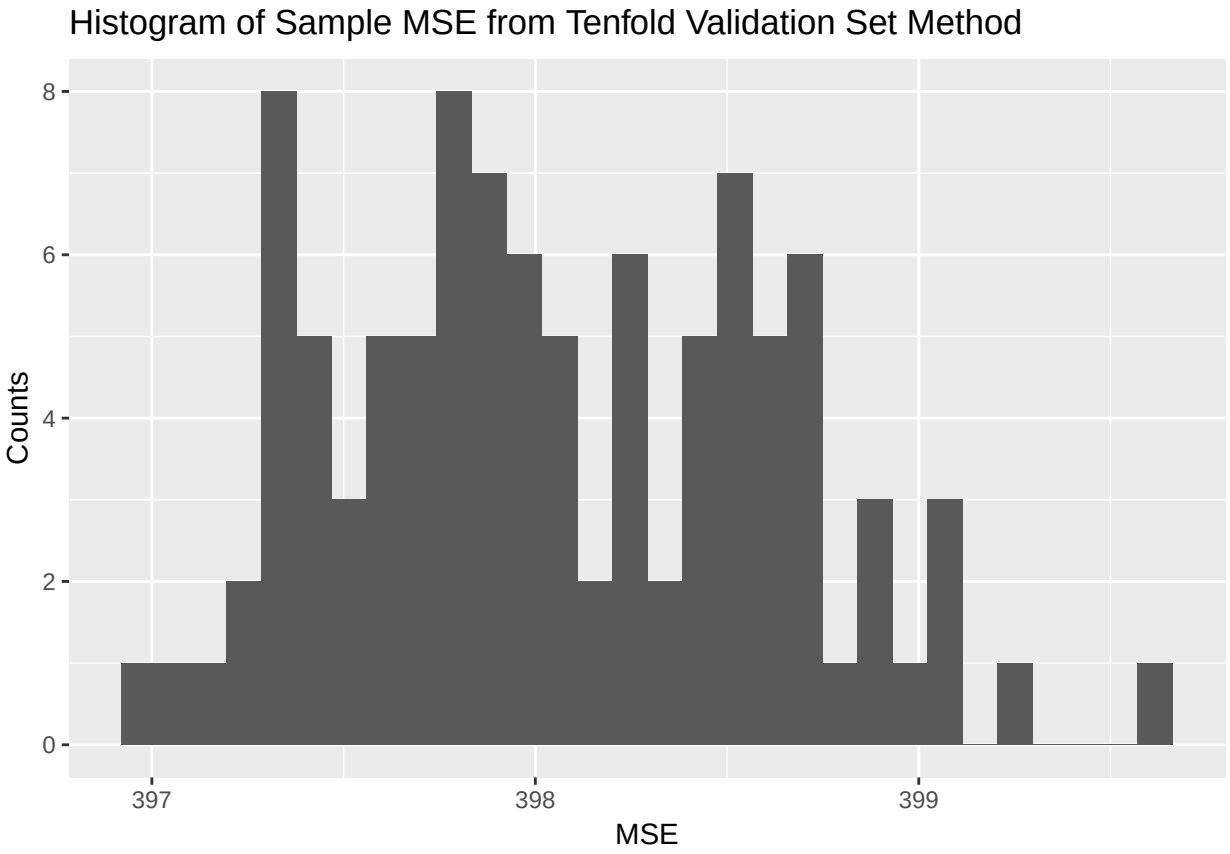
```
## [1] "1 f: The mean MSE for 10 fold cross validation approach simulating 100 times is:  398.064164615
```

```
## [1] 0.5602606
```

The mean squared error for 10 fold cross validation simulating 100 times is 398.06. Again this result doesn't differ from the full sample method much. And the mean squared error is computed for the test data on the training dataset. Though the MSE performance doesn't improve compared to the validation set method, the

standard deviation of the MSE among the 100 ten fold validation sets are better than that of the validation set method. In the validation set method, the standard deviation of the MSE is higher(8.83) compared to the ten fold cross validation method(0.57). This corroborates the idea that ten fold cross validation and LOOCV method has less fluctuation in MSE compared to the validation set method.

The following is the histogram of the tenfold cross validation approach. As we can see, compared to the validation set approach, the range of the sample MSE is much smaller, ranging from (397, 399). While the sample MSE of validation set approach ranges from (370, 420).



7.

Table 3: Estimated Results for Bootstrap

	term	est.boot	se.boot
1	(Intercept)	58.96	2.95
2	age	0.05	0.03
3	dem	15.43	1.11
4	educ	-0.35	0.19
5	female	4.08	0.95
6	rep	-15.89	1.43

From the result we can see that the bootstrap result is similar to that of the simple linear regression in a). But we get a larger standard error for each estimator. This comes as a result of that the assumption of OLS is not fully satisfied.

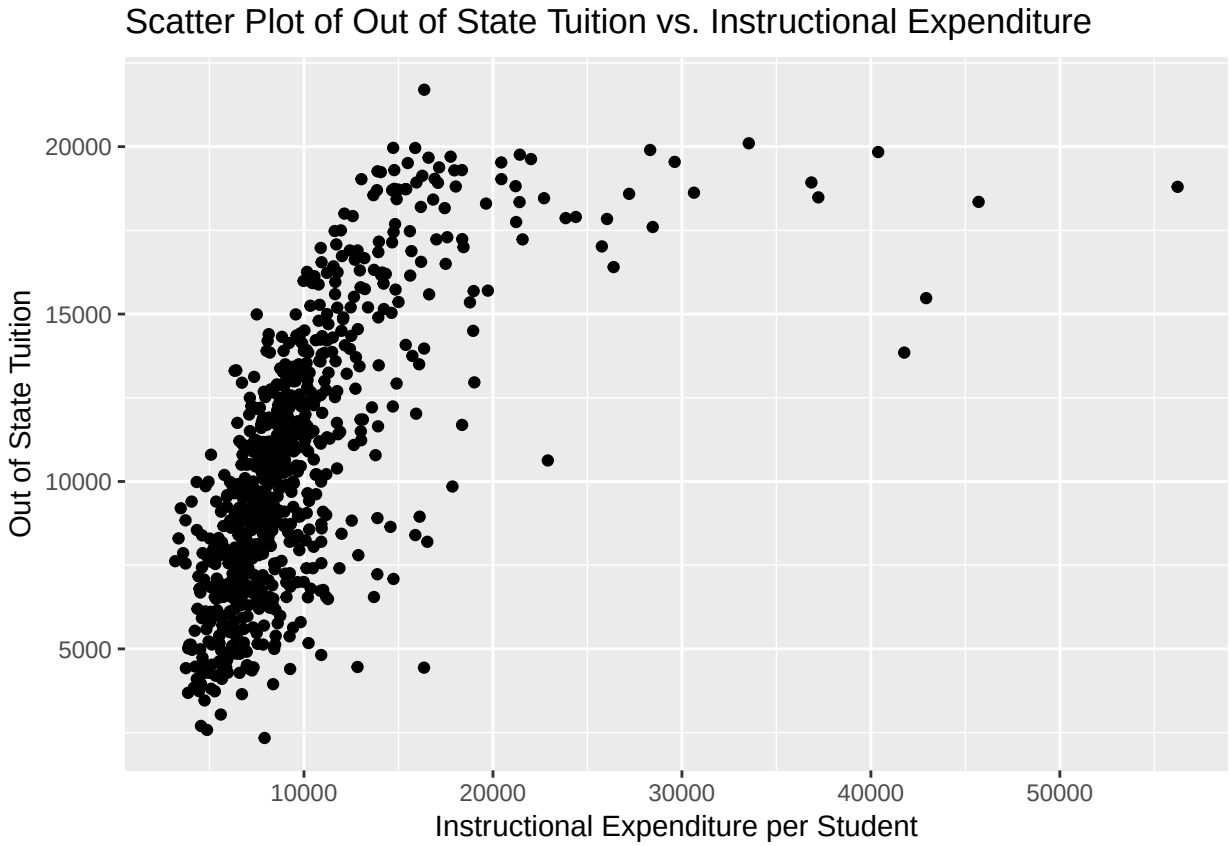
Problem 2

We first estimate the relationship between outstate tuition and the school type. For our sample set, we have that the average tuition for public school is 6813.41 dollars. The estimated coefficient is 4988.283 dollars. This means that compared to public university, the tuition for private university is on average 4988.23 dollars higher than private school (5/7 time higher than that of pulic school). This effect is relatively large.

Table 4: Linear Regression Outstate Tuition vs. Private School

	Dependent variable:
	Outstate
PrivateYes	4,988.283***
	(270.216)
Constant	6,813.410***
	(230.422)
Observations	777
R^2	0.305
Adjusted R^2	0.305
Residual Std. Error	3,354.999 (df = 775)
F Statistic	340.785*** (df = 1; 775)
Note:	*p<0.1; **p<0.05; ***p<0.01

Now we can look at the relationship between out of state tuition and instructional expenditure per student. We first plot the scatter plot. This plot is not linear by eyeballing the scatter plot. Thus we can estiamte the tuition elasticity of instructional expenditure (i.e. take log of both response variables and predictors).
We summarize both of the models in Table 5. For the level model, one dollar increase in instructional expenditure leads to 0.518 dollar increase in out of state tuition. For the log model. We can see that one percent increase in instructional expenditure leads to 0.706 percent increase in out of state tuition. Firstly examining the R square value, we can see that there is a 0.05 increase in R square. But if we take log on only the instructional expenditure, we can further improve R square value to 0.577. This model indicates that one percent in instructional expenditure leads to 7482.15 dollars increase in out of state tuition.
Then we can further compare both models by using validation set method.
The following is the result.



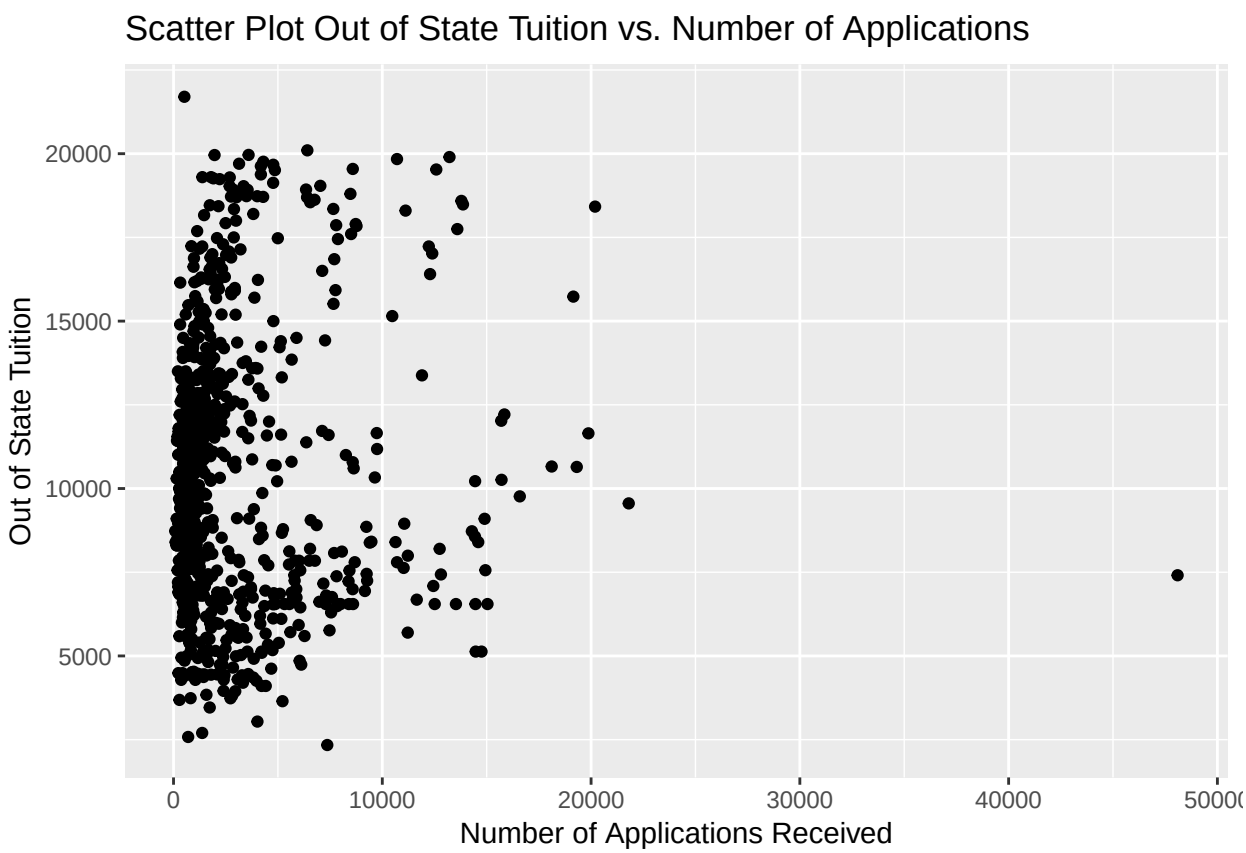

```
## [1] "MSE for the Simple Linear Regression without Log Transformation:"
## [1] 9108730
## [1] "MSE for the Simple Linear Regression WITH Log Transformation on Both Sides:"
## [1] 7518781
## [1] "MSE for the Simple Linear Regression WITH Log Transformation on Predictor:"
## [1] 7089590
```

Above is the MSE for the test set for the three model. We can see that the one side log transformation performs the best. Thus judging by the regression result, university in the US calculate their budget for out of state tuition in a way that they project the percent change of instructional expenditure into a level change in out of state tuition.

From a demand and supply side perspective, does higher demand for the education for one school increases the tuition for one school? Thus we can examine the relatinoship between the out of state tuition and number of application received each year. Firstly we plot the scatter plot:

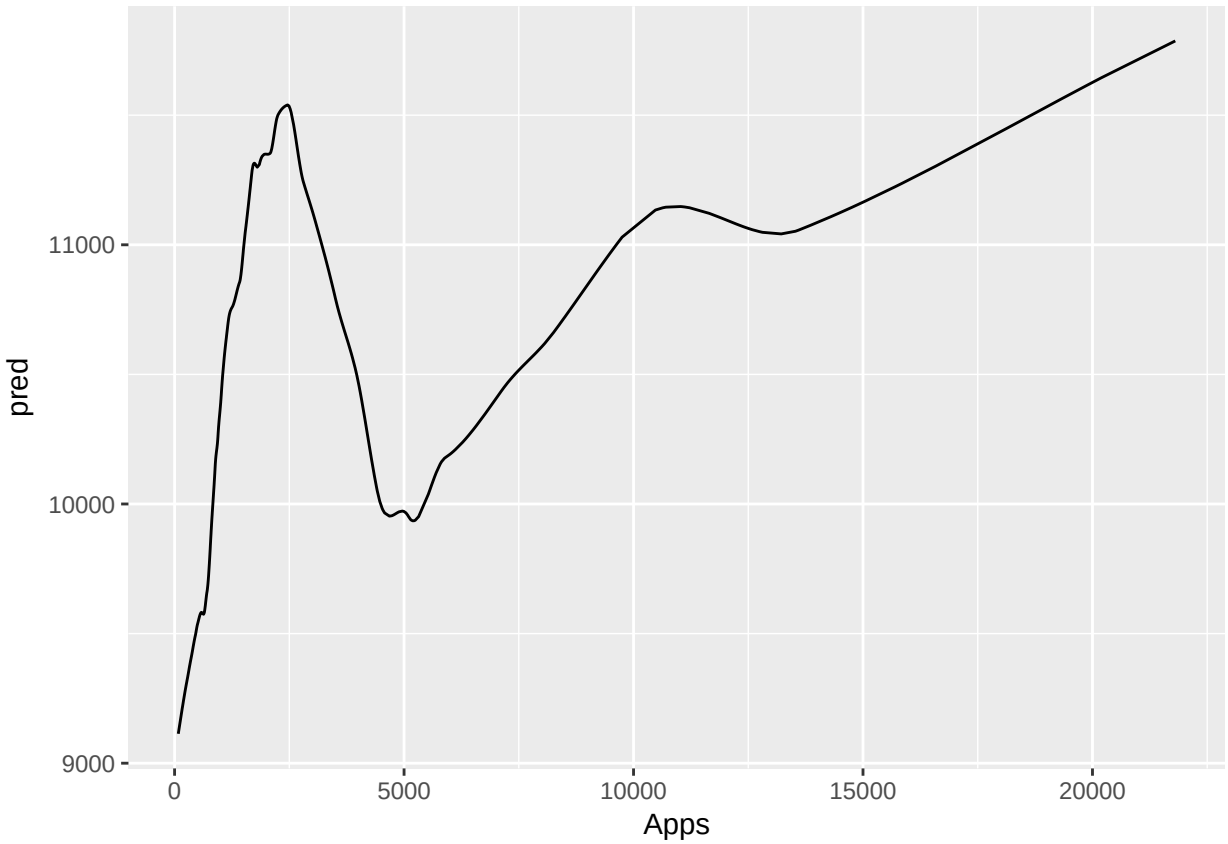
Table 5: Out of State Tuition vs. Instructional Expenditure (Level vs. Log)

	Dependent variable:		
	Outstate	log(Outstate)	Outstate
	(1)	(2)	(3)
Expend	0.518***		
	(0.020)		
log(Expend)		0.706***	7,482.150***
		(0.025)	(229.915)
Constant	5,433.512***	2.767***	−57,502.040***
	(224.806)	(0.227)	(2,089.888)
Observations	777	777	777
R^2	0.453	0.507	0.577
Adjusted R^2	0.452	0.506	0.577
Residual Std. Error (df = 775)	2,978.324	0.285	2,616.838
F Statistic (df = 1; 775)	640.864***	796.386***	1,059.051***
Note:			*p<0.1; **p<0.05; ***p<0.01



By looking at the scatter plot, we can see that there is different trend for application reveived among different regions. Specifically, when number of applications received below and above 5000 dollars, there are different trends. But since the scatter plot is noisy, we can fit a local regression to test this relationship.

By using the full sample to test the local regression, we can plot the following prediction plot. ([Note] in this case, we exclude the outlier; we do this by limiting the number of applications below 30000.)



By looking at the graph of the local regression, we can see that there is different trend below and above 5000 applications.
When nthe number of applications is below 5000, we see a hump shaped curve. This may indicate that there are some colleges that will allow paying to get into college so that they can take advantage of the high tuition to generate profit. These are usually business type colleges. But it's still hard for us to explain the drop in the region between 2500 and 5000. One possible explanation is that we enter a region of academic institutes and there is a break around 2500. After 5000 applications, we can see a normal supply and price curve. Though for academic institute, a demand and supply analysis may oversimply the analysis by excluding externality and government subsidies, by looking at the graph, there is still an element of demand and supply relationship.

```
## [1] "MSE for Local Regression using tenfold cross validation: "
```

```
## [1] 15620497
```

```
## [1] "MSE for OLS using tenfold cross validation: "
```

```
## [1] 16100881
```

The above is the MSE for OLS and Local Regression. As we can see local regression does improve the model perfermance and reduce MSE.

In summary, we find that the out of state tuition for private university is almost twice the amount of public school. There is a linear-log relationship between out of state tuition and instructinoal expenditure. This indicates that the university projects the percent change of instructional expenditure into a level change of out of state tuition. And finally, there is a supply side effect on out of state tuition for the schools that have out of state tuition higher than 5000 dollars but the effect is reversed for the schools between 2500 and 5000 dollars. This may indicate discontinuity in the sample.

3.

1.

2.

Below is the regression result:

Table 6: Simple Linear Regression of Out of State Tuition

	Dependent variable:
	Outstate
PrivateYes	2,879.279***
	(253.693)
Room.Board	1.042***
	(0.100)
PhD	37.783***
	(6.922)
perc.alumni	50.772***
	(8.797)
Expend	0.209***
	(0.021)
Grad.Rate	24.117***
	(6.371)
Constant	−3,637.358***
	(520.491)
Observations	544
R^2	0.751
Adjusted R^2	0.749
Residual Std. Error	2,036.043 (df = 537)
F Statistic	270.404*** (df = 6; 537)
Note:	*p<0.1; **p<0.05; ***p<0.01

The model has an R square of 0.751. This means that approximately 75.1% of the variation in out of state tuition is explained by the response variables chosen. This is considered a good fit.

Additionally, all of the chosen variables are significant. The regression shows that holding all other constant, on average the out of state tuitoin is 2879.279 dollars higher for private university than for public university. On average, one dollar increase in room and board expenses will indicate 1.042 dollar increase in tuition. Also, one percent increase in the number of faculties who have PhD degree will on average lead to 37.783 dollars increase in out of state tuition. And one percent increase in Alumni donation will also on average increase out of state tuition by 50.772 dollars.

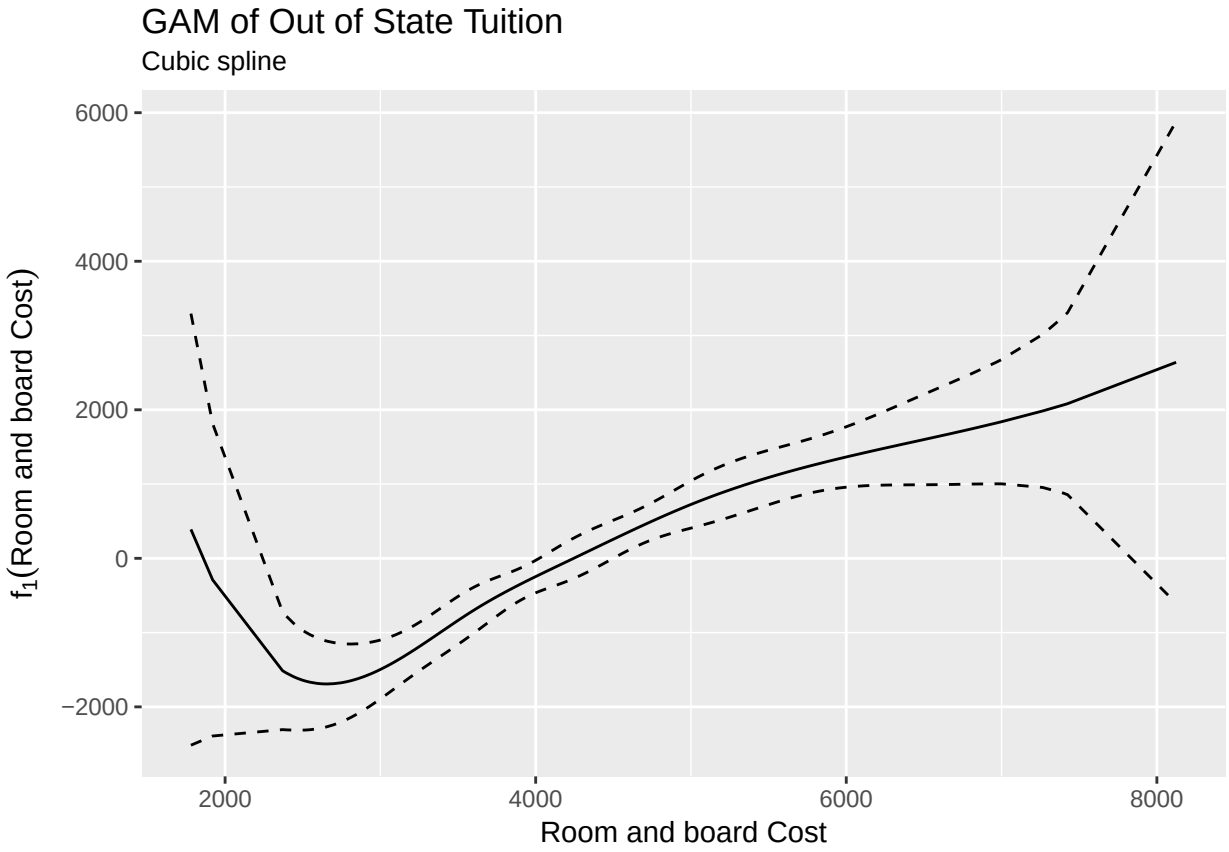
A somewhat odd result is that one dollar increase in instructional expenditure leads to a 0.209 dollar increase on average in tuition. This is even smaller compared to the effect of room and board cost because normally the tuition doesn't include room and board cost but does include instructional expenditure cost. One plausible explanation is indicated in the above question. The university change the level of tuition in response to the percent change of instructional expenditure. A one percent increase in instructinoal expenditure will increase dollar amount change of tuition.

Grauduation rate is also significant. One percent increase in graduation rate will lead to 24.117 dollar increase in tuition. This variable can have several possible effects. One effect is through the resources the university spend on student education: for example, more tutors, higher teacher sallary etc. In general, more resources

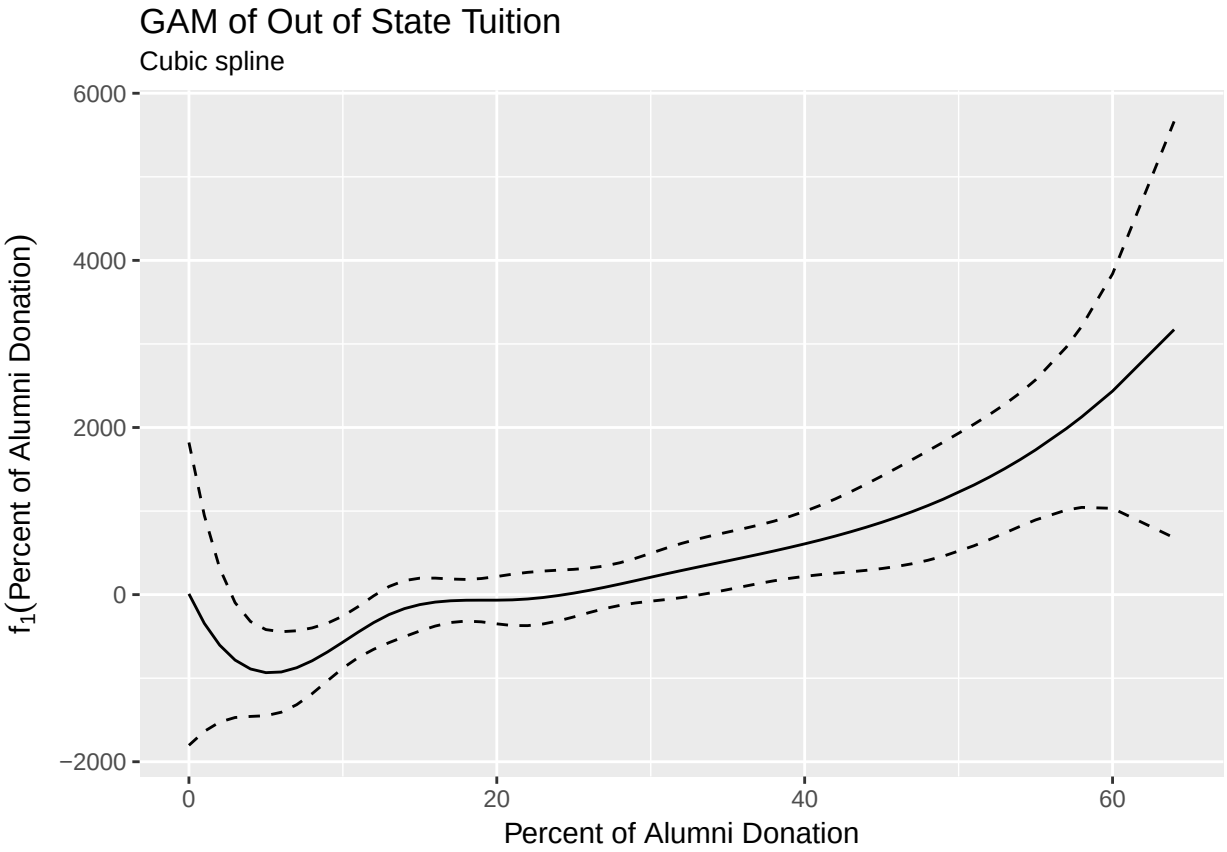
spent on student education, higher the chance for the students to graduate. Thus, this variable can also have another effect through reputation. The higher quality one school give, the higher the reputation they gain. And it's normal for the Ivy league university to have the highest tuition among all of univeristy.

3.

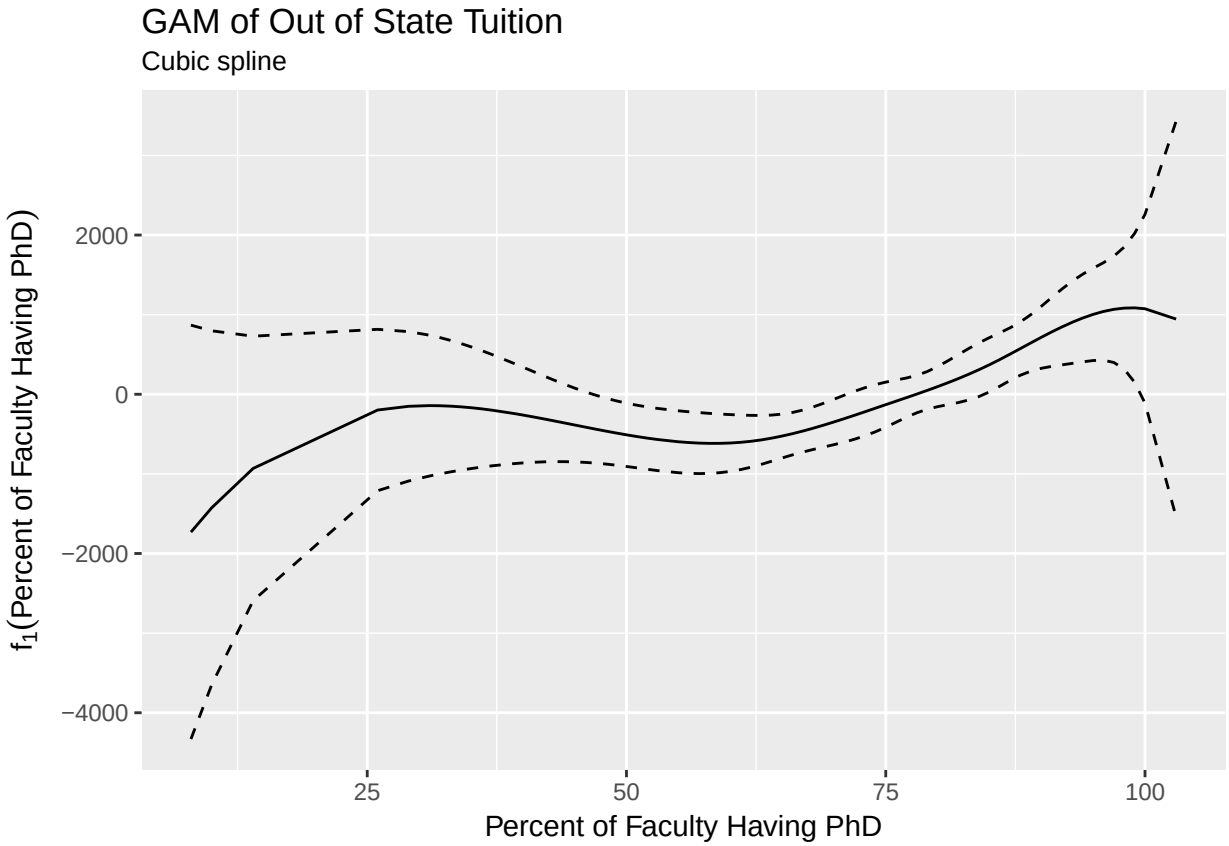
```
##
## Call: gam(formula = Outstate ~ Private + bs(Room.Board, df = 6) + bs(perc.alumni,
##     df = 6) + bs(PhD, df = 6) + log(Expend) + lo(Grad.Rate),
##     family = gaussian, data = college_split$train)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -7457.6 -1223.3   -36.4  1251.2  8727.4
##
## (Dispersion Parameter for gaussian family taken to be 3699546)
##
##     Null Deviance: 8951828671 on 543 degrees of freedom
## Residual Deviance: 1921734869 on 519.4516 degrees of freedom
## AIC: 9797.086
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private                   1.00 2807121275 2807121275 758.775 < 2.2e-16 ***
## bs(Room.Board, df = 6)    6.00 2233164469  372194078 100.605 < 2.2e-16 ***
## bs(perc.alumni, df = 6)   6.00  827007160  137834527  37.257 < 2.2e-16 ***
## bs(PhD, df = 6)           6.00  626312096  104385349  28.216 < 2.2e-16 ***
## log(Expend)               1.00  435297973  435297973 117.662 < 2.2e-16 ***
## lo(Grad.Rate)             1.00   70972692   70972692  19.184 1.436e-05 ***
## Residuals               519.45 1921734869    3699546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                           Npar Df Npar F   Pr(F)
## (Intercept)
## Private
## bs(Room.Board, df = 6)
## bs(perc.alumni, df = 6)
## bs(PhD, df = 6)
## log(Expend)
## lo(Grad.Rate)                2.5  2.319 0.08495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



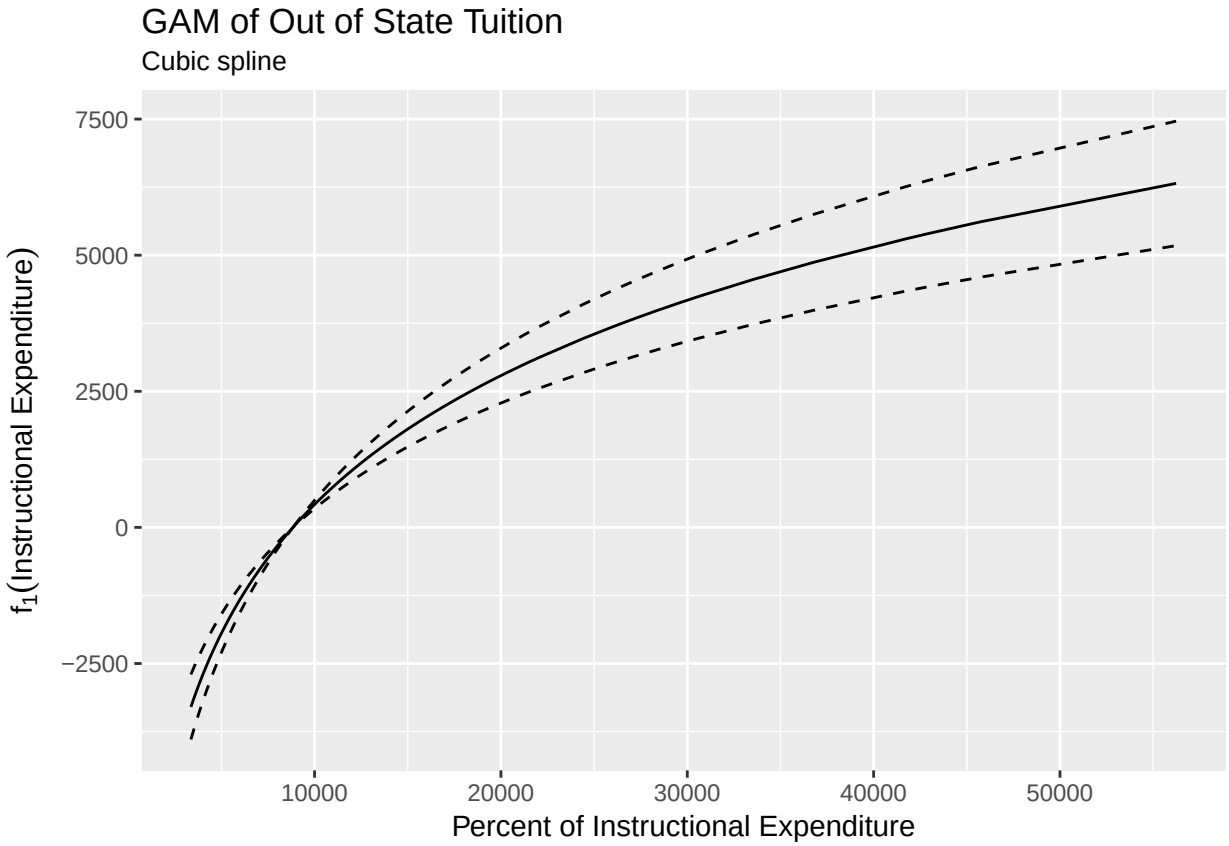
The above graph is the relationship between room and board cost and out of state tuition. I fit a spline regression with 3 knots. In general there is a positive relationship between these two variables. The curvature only occurs at the end of the range. Thus we can't conclude there is much non linear relationship between these two variables.



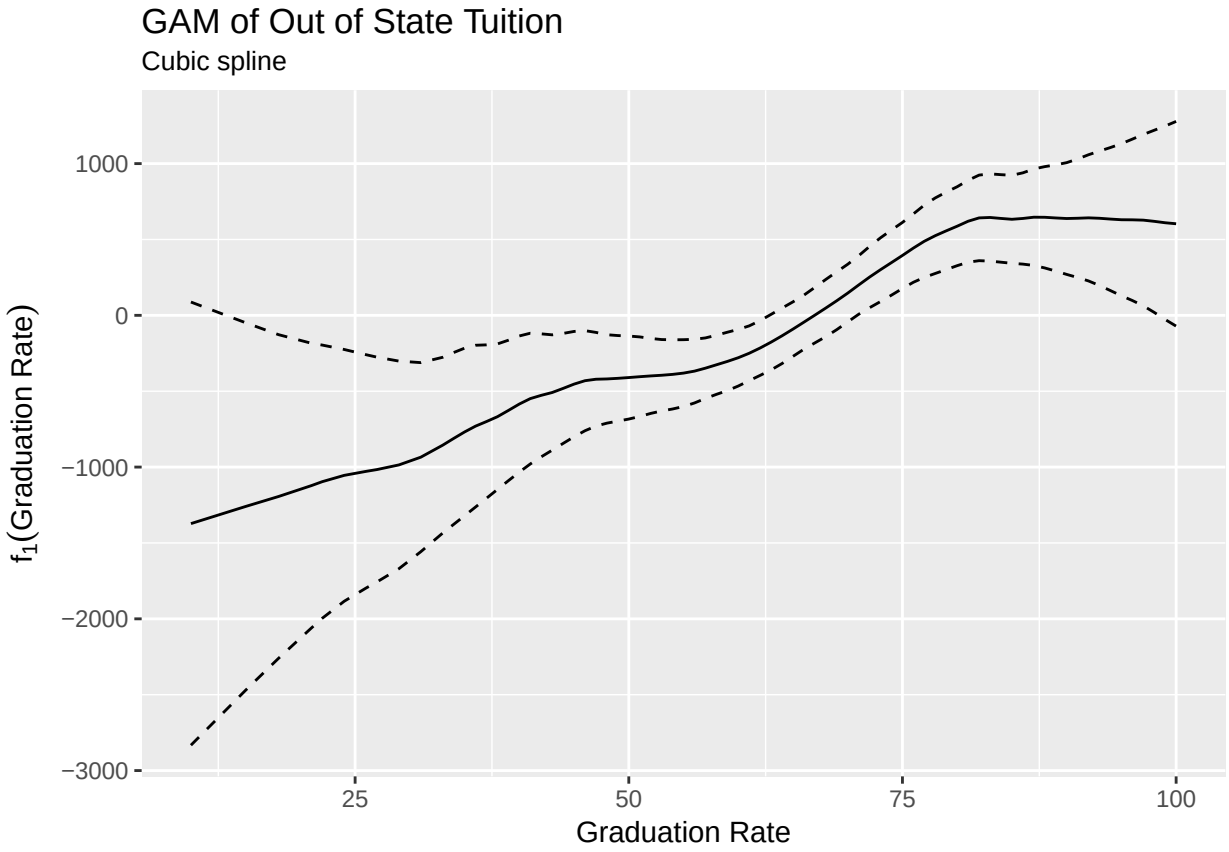
The above graph is the basis function for percent of alumni who donates. I fit a spline regression with 3 knots. The line demonstrates a slowly increasing trend. There is curvature on the end of the range. This is caused by the relatively sparse points at each end of the range. We can't conclude that there is much non linear relationship between out of state tuition and percent of alumni who donates.



The above grpah is the relationship between Percent of Faculty having PhD and out of state tuition. I fit a spline regression with 3 knots. We can see that the grpah demonstrates a relatively flat line with curvature only at the end of the graph. This indicates there is relatively weak relationship between these two varialbles.



The above is the relationship between out of state tuition and instructional expenditure. I fit a log transformation of the variable into the model. This decision is made after comparing the performance of a local regression spline regression. Log transformation gives the best fit in the sense that it has narrower confidence interval. This again confirms our conclusion above that the university projects percent change of instructional expenditure to level change in level of out of state tuition.



The above is the relationship between out of state tuition and graduation rate. Since by looking at the scatter plot of these two variables, the relationship is unclear, I fit a local regression. The above graph shows a lienar trend. But the relationship is not string since the confidence interval band is really wide. This indicates there is not much relationship between out of state tuition and graduation rate. Hence this also rejects our hypothesis that graduation rate can affect tuition trhough quality of education and reputation.

4.

```
## [1] "MSE for OLS model is:  4109016.08528828"
```

```
## [1] "MSE for OLS model is:  4006499.03929072"
```

As we can see, the MSE for GAM model is better than OLS model. This may indicate the non-linear model is better. But note that this may come as a result of overfitting the model. Apparently, transforming 5 variables into non-linear basis functions decreases MSE by approximately 1/40 is up for debate whether it's good enough. Combine the graph generated above, the non-linear model may acutally overfit the model since most of the variables exhibits linear trend.

5

```
## Analysis of Variance Table
##
## Model 1: Outstate ~ Private + Room.Board + PhD + perc.alumni + Expend +
##     Grad.Rate
## Model 2: Outstate ~ Private + bs(Room.Board, df = 6) + PhD + perc.alumni +
##     Expend + Grad.Rate
```

```
## Model 3: Outstate ~ Private + bs(Room.Board, df = 6) + bs(PhD, df = 6) +
##     perc.alumni + Expend + Grad.Rate
## Model 4: Outstate ~ Private + bs(Room.Board, df = 6) + bs(PhD, df = 6) +
##     bs(perc.alumni, df = 6) + Expend + Grad.Rate
## Model 5: Outstate ~ Private + bs(Room.Board, df = 6) + bs(PhD, df = 6) +
##     bs(perc.alumni, df = 6) + log(Expend) + Grad.Rate
## Model 6: Outstate ~ Private + bs(Room.Board, df = 6) + bs(PhD, df = 6) +
##     bs(perc.alumni, df = 6) + log(Expend) + lo(Grad.Rate)
## Model 7: Outstate ~ Private + bs(Room.Board, df = 6) + bs(perc.alumni,
##     df = 6) + bs(PhD, df = 6) + log(Expend) + lo(Grad.Rate)
##   Res.Df        RSS     Df Sum of Sq      F   Pr(>F)
## 1 537.00 2226118933
## 2 532.00 2181346587 5.0000  44772347 2.4204 0.034819 *
## 3 527.00 2107375817 5.0000  73970770 3.9989 0.001437 **
## 4 522.00 2088489699 5.0000  18886117 1.0210 0.404432
## 5 522.00 1943598550 0.0000 144891150
## 6 522.00 1943598550 0.0000         0
## 7 519.45 1921734869 2.5484  21863680 2.3190 0.084947 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above is the anova table for the GAM models and OLS model. We transform the linear predictor into a non-linear basis functions one by one. Then by compare each consecutive model, we can know which variable has non-linear relationship with the out of state tuition.

I don't think there is a strong evidence that a non-linear specification is necessary judging by the grpah and the anova test.

The result indicates that room and board cost and percent of faculties having PhD degree are the two variables that have non-linear relationship with out of state tuition. All of the other variables do not exhibit strong non-linear relationship with out of state tuition.

The result goes against our observation of the relatively flat regression line between percent of faculties having PhD degree and out of state tuition. Faculties having PhD degree exhibits different rate of change in a certain degree: at each end of the range, the rate of out of state tuition increases faster with respect to percent of faculties having PhD degree. The rate of change is smallest in the middle. But again, this may be a result of the broad confidence interval at the end of the range where observation points are sparse.

But in general, I don't think there is a strong evidence of non-linearity between the response variable and the predictors.