Problem set #7: resampling and nonlinearity

Weijia Li

Table of Contents

Part 1: Sexy Joe Biden (redux) [4 points] 1
Part 2: College (bivariate) [3 points] 5
Room.Board 5
Accept 8
Expend 11
Part 3: College (GAM) [3 points] 14

Part 1: Sexy Joe Biden (redux) [4 points]

For this exercise we consider the following functional form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

where Y is the Joe Biden feeling thermometer, X_1 is age, X_2 is gender, X_3 is education, X_4 is Democrat, and X_5 is Republican.[1] Report the parameters and standard errors.

1. Estimate the training MSE of the model using the traditional approach.
   - Fit the linear regression model using the entire dataset and calculate the mean squared error for the training set.

```
##          term     estimate std.error  statistic      p.value
## 1 (Intercept)  58.81125899 3.1244366  18.822996 2.694143e-72
## 2         age   0.04825892 0.0282474   1.708438 8.772744e-02
## 3      female   4.10323009 0.9482286   4.327258 1.592601e-05
## 4        educ  -0.34533479 0.1947796  -1.772952 7.640571e-02
## 5         dem  15.42425563 1.0680327  14.441745 8.144928e-45
## 6         rep -15.84950614 1.3113624 -12.086290 2.157309e-32

##
## Call:
## lm(formula = biden ~ age + female + educ + dem + rep, data = biden)
##
## Residuals:
```

[1] Independents must be left out to serve as the baseline category, otherwise we would encounter perfect multicollinearity.

```
##     Min      1Q  Median      3Q     Max
## -75.546 -11.295   1.018  12.776  53.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.81126    3.12444  18.823  < 2e-16 ***
## age           0.04826    0.02825   1.708   0.0877 .
## female        4.10323    0.94823   4.327 1.59e-05 ***
## educ         -0.34533    0.19478  -1.773   0.0764 .
## dem          15.42426    1.06803  14.442  < 2e-16 ***
## rep         -15.84951    1.31136 -12.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.91 on 1801 degrees of freedom
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2795
## F-statistic: 141.1 on 5 and 1801 DF,  p-value: < 2.2e-16
```

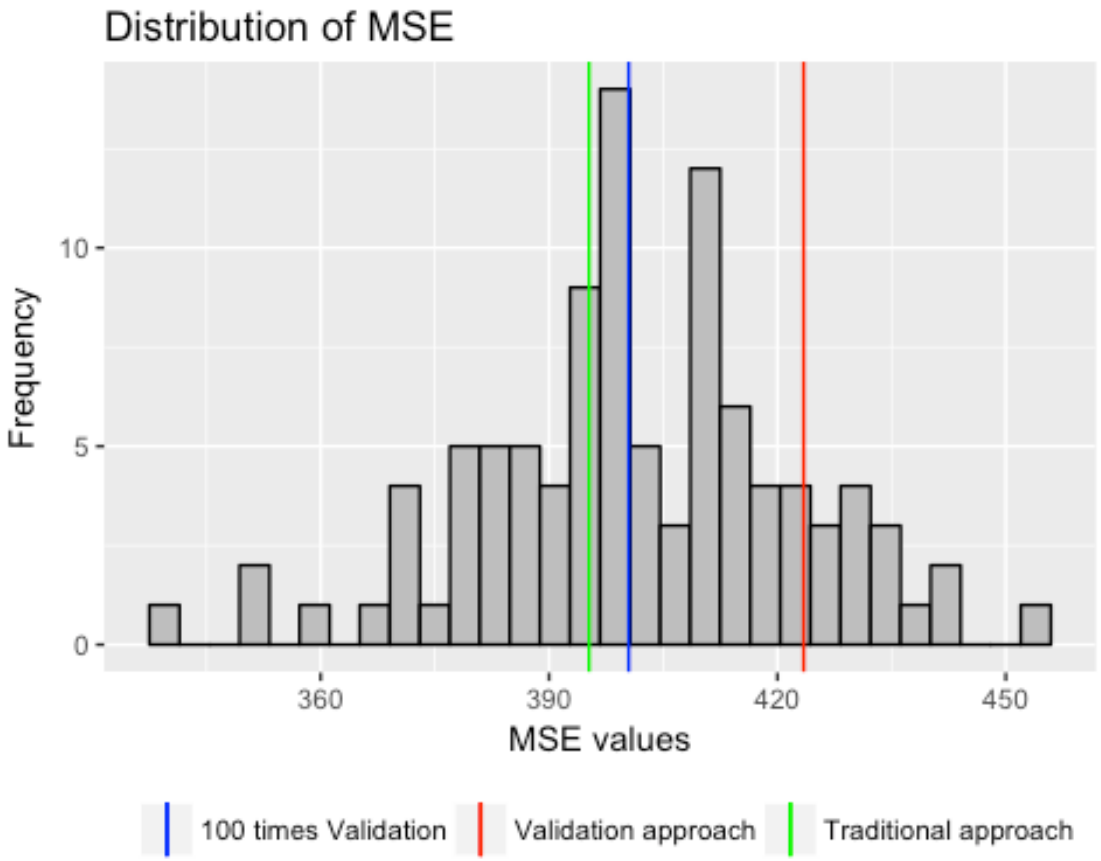
The MSE using traditional approach is 395.2702.

2. Estimate the test MSE of the model using the validation set approach.

The MSE calculated only on the test set is 399.8303, which is slightly larger than the traitional approach.

3. Repeat the validation set approach 100 times, using 100 different splits of the observations into a training set and a validation set. Comment on the results obtained.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   340.9   390.2   400.1   401.7   414.6   455.3
```



From the histogram above, MSE is most frequently at the mean(401.7). This is larger, but only slightly, than one time validation.

4. Estimate the test MSE of the model using the leave-one-out cross-validation (LOOCV) approach. Comment on the results obtained.

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

The mean MSE value using LOOCV approach is 397.9555. This is smaller than the value from the 100-times simulation but larger than the value from traditional approach.

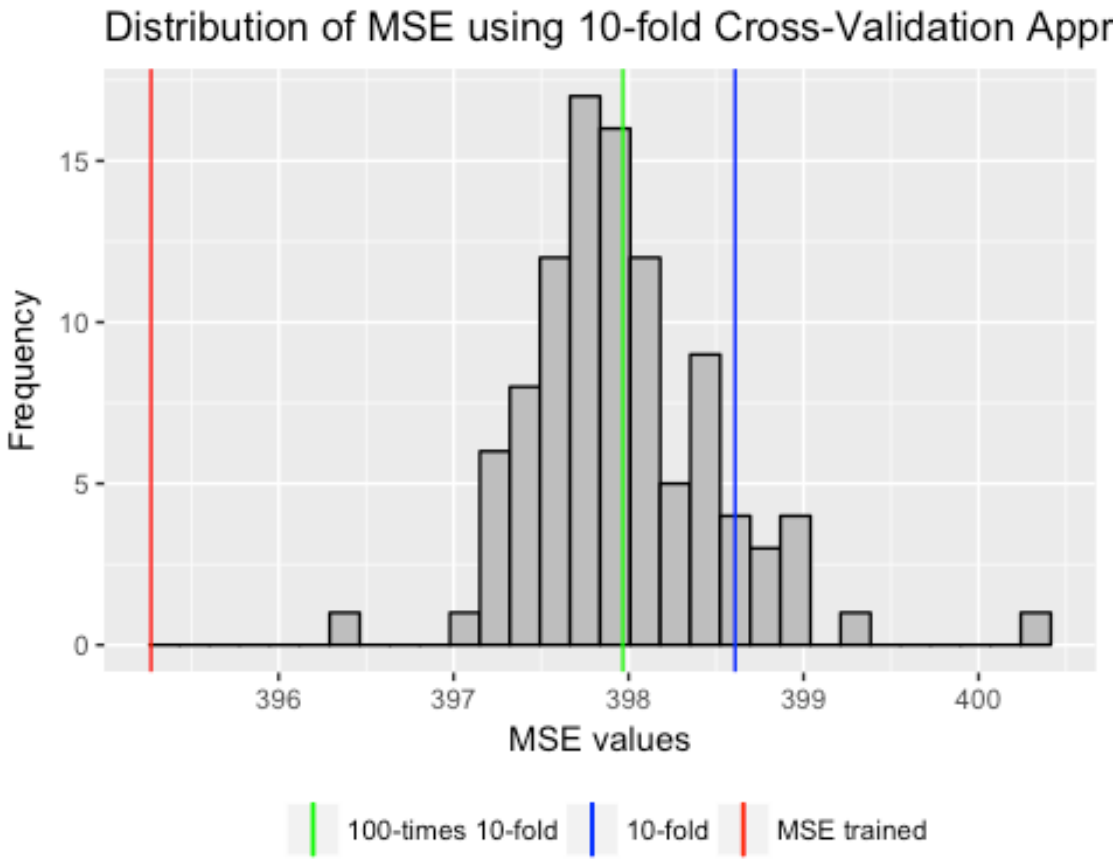
5. Estimate the test MSE of the model using the 10-fold cross-validation approach. Comment on the results obtained.

```
## [1] 397.414
```

Now the MSE is 398.0729, which is very close to what I got in LOOCV approach and, again, slightly higher than the traditional approach for the training set.

6. Repeat the 10-fold cross-validation approach 100 times, using 100 different splits of the observations into 10-folds. Comment on the results obtained.

```
## [1] 397.9667
```



By plotting the histogram we can see the gap between the result in part 1 is significant. Most of the results are between 397 and 399 and the mean is 398.178.

7. Compare the estimated parameters and standard errors from the original model in step 1 (the model estimated using all of the available data) to parameters and standard errors estimated using the bootstrap ($n = 1000$).

```
## # A tibble: 6 × 3
##          term     est.boot   se.boot
##         <chr>        <dbl>     <dbl>
## 1 (Intercept)  58.91521173 2.9785509
## 2         age   0.04769957 0.0288367
## 3         dem  15.42937778 1.1081598
## 4        educ  -0.34943659 0.1921614
## 5      female   4.08557875 0.9503575
## 6         rep -15.87451623 1.4442938

##          term     estimate std.error  statistic      p.value
## 1 (Intercept)  58.81125899 3.1244366  18.822996 2.694143e-72
## 2         age   0.04825892 0.0282474   1.708438 8.772744e-02
## 3      female   4.10323009 0.9482286   4.327258 1.592601e-05
## 4        educ  -0.34533479 0.1947796  -1.772952 7.640571e-02
## 5         dem  15.42425563 1.0680327  14.441745 8.144928e-45
## 6         rep -15.84950614 1.3113624 -12.086290 2.157309e-32
```

The bootstrap standard error and the standard error generated in part 1 are about the same but ther results from boostrap are mostly slightly larger (apart from standard error in 'dem') and the estimates also vary little. I am expecting the bootstrap standard errors to be larger than the other one since they don't depend on distributional assumptions.

Part 2: College (bivariate) [3 points]

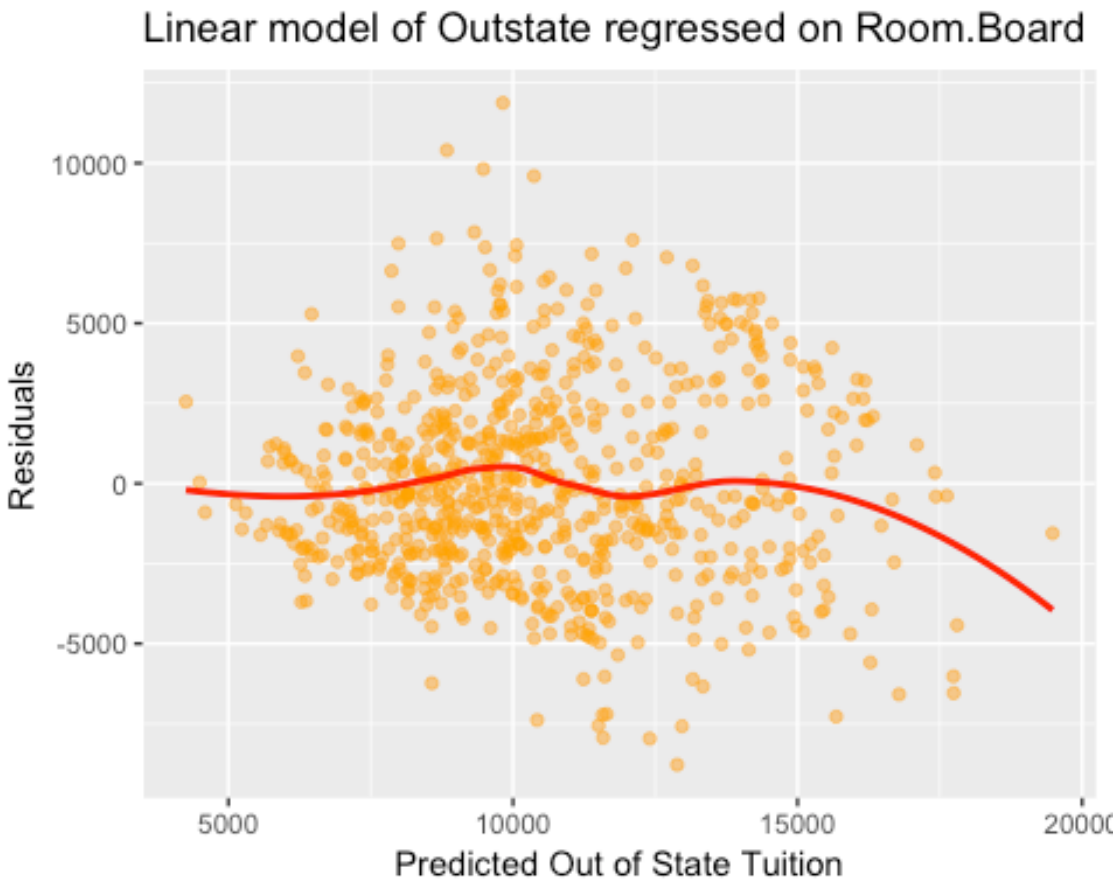
```
##
## Call:
## lm(formula = Outstate ~ ., data = c_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6782.6 -1267.5   -40.9  1244.5  9953.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.587e+03  7.660e+02  -2.072  0.03860 *
## PrivateYes   2.264e+03  2.480e+02   9.128  < 2e-16 ***
## Apps        -3.034e-01  6.734e-02  -4.506 7.64e-06 ***
## Accept       8.124e-01  1.293e-01   6.286 5.51e-10 ***
## Enroll      -5.492e-01  3.541e-01  -1.551  0.12134
## Top10perc    2.834e+01  1.098e+01   2.582  0.01002 *
## Top25perc   -3.779e+00  8.475e+00  -0.446  0.65576
## F.Undergrad -9.567e-02  6.152e-02  -1.555  0.12038
## P.Undergrad  1.166e-02  6.049e-02   0.193  0.84720
## Room.Board   8.816e-01  8.558e-02  10.302  < 2e-16 ***
## Books       -4.592e-01  4.479e-01  -1.025  0.30551
## Personal    -2.294e-01  1.183e-01  -1.940  0.05280 .
## PhD          1.124e+01  8.730e+00   1.288  0.19822
## Terminal     2.467e+01  9.538e+00   2.587  0.00988 **
## S.F.Ratio   -4.644e+01  2.441e+01  -1.902  0.05753 .
## perc.alumni  4.180e+01  7.561e+00   5.528 4.45e-08 ***
## Expend       1.990e-01  2.269e-02   8.769  < 2e-16 ***
## Grad.Rate    2.400e+01  5.506e+00   4.359 1.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1958 on 759 degrees of freedom
## Multiple R-squared:  0.7684, Adjusted R-squared:  0.7632
## F-statistic: 148.1 on 17 and 759 DF,  p-value: < 2.2e-16
```

Choosing the most significant variables to do a linear fit.

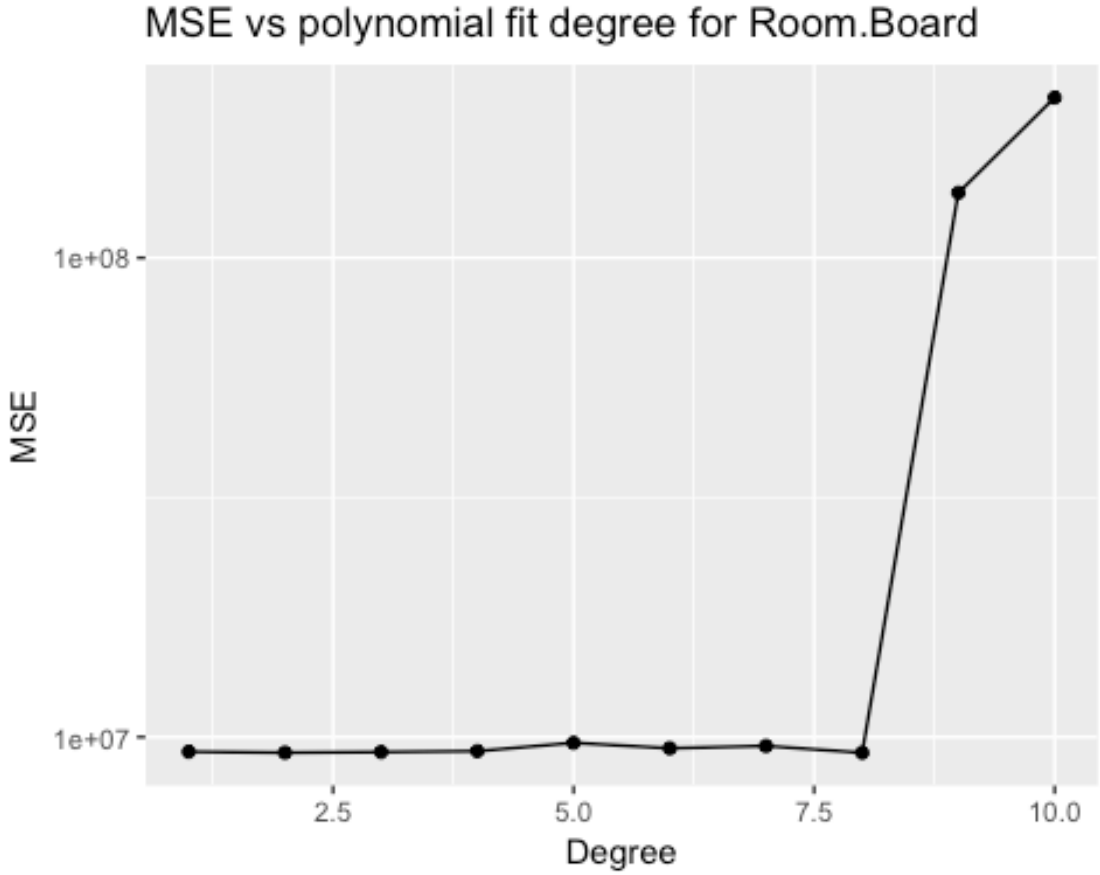
Room.Board

```
##
## Call:
## lm(formula = Outstate ~ Room.Board, data = c_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8781.0 -2070.6  -350.8  1877.4 11877.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.44525  447.76786  -0.039    0.969
## Room.Board    2.40001    0.09965  24.084   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3044 on 775 degrees of freedom
## Multiple R-squared:  0.4281, Adjusted R-squared:  0.4273
## F-statistic:   580 on 1 and 775 DF,  p-value: < 2.2e-16
```

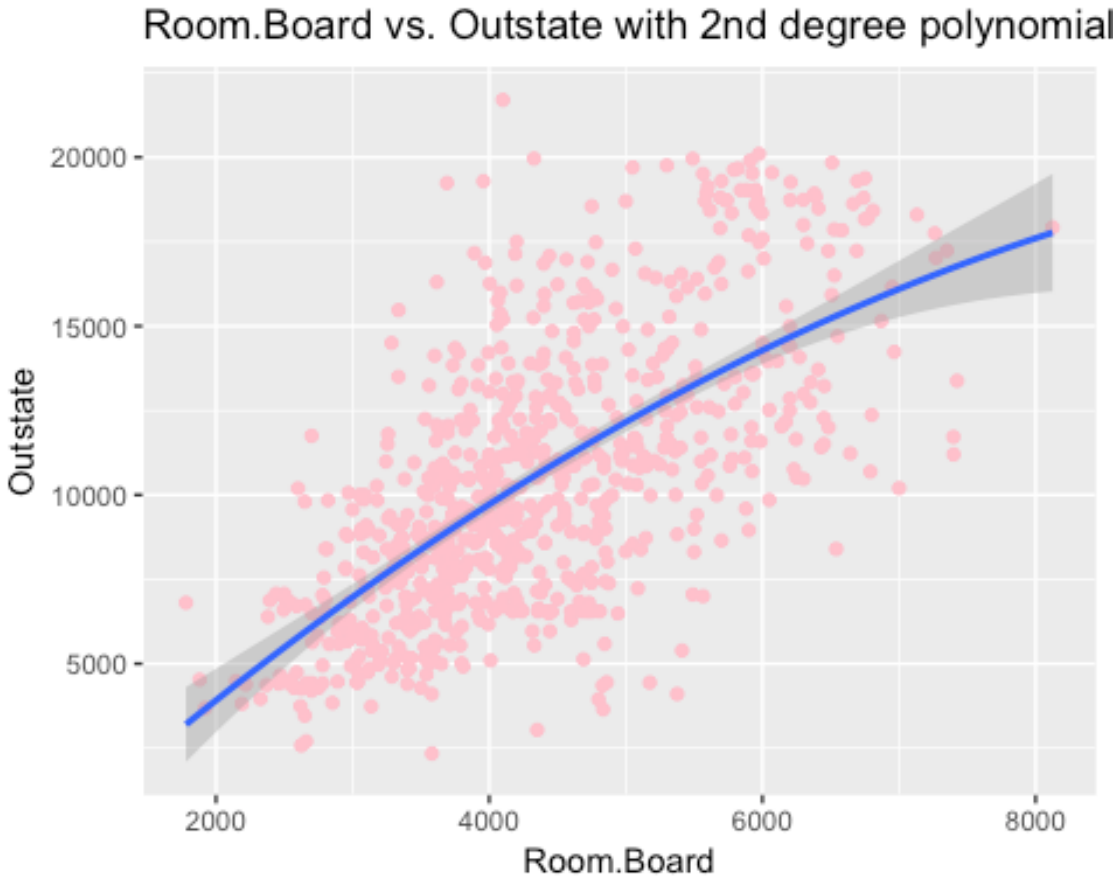


The relationship is indeed not linear, yet a higher degree polynomials may have a better



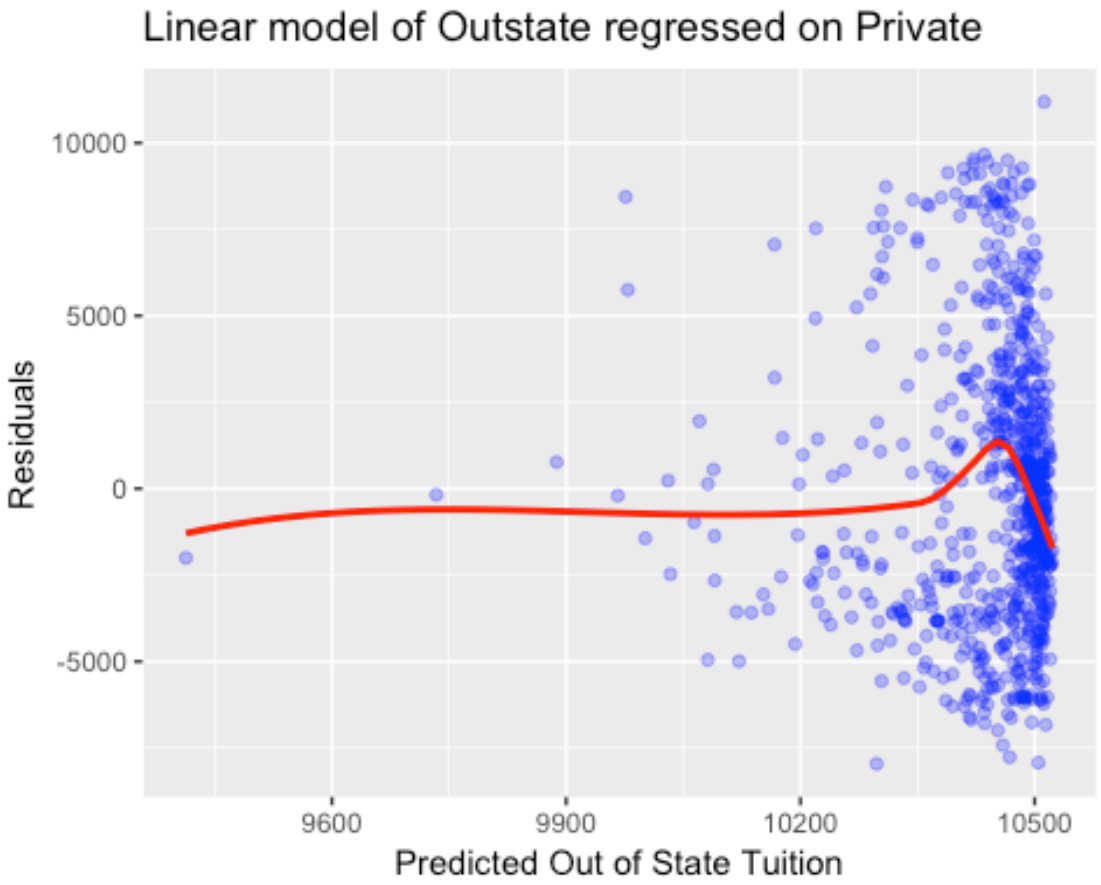
result.

From the plot, a second degree polynomial would be enough for linearity.

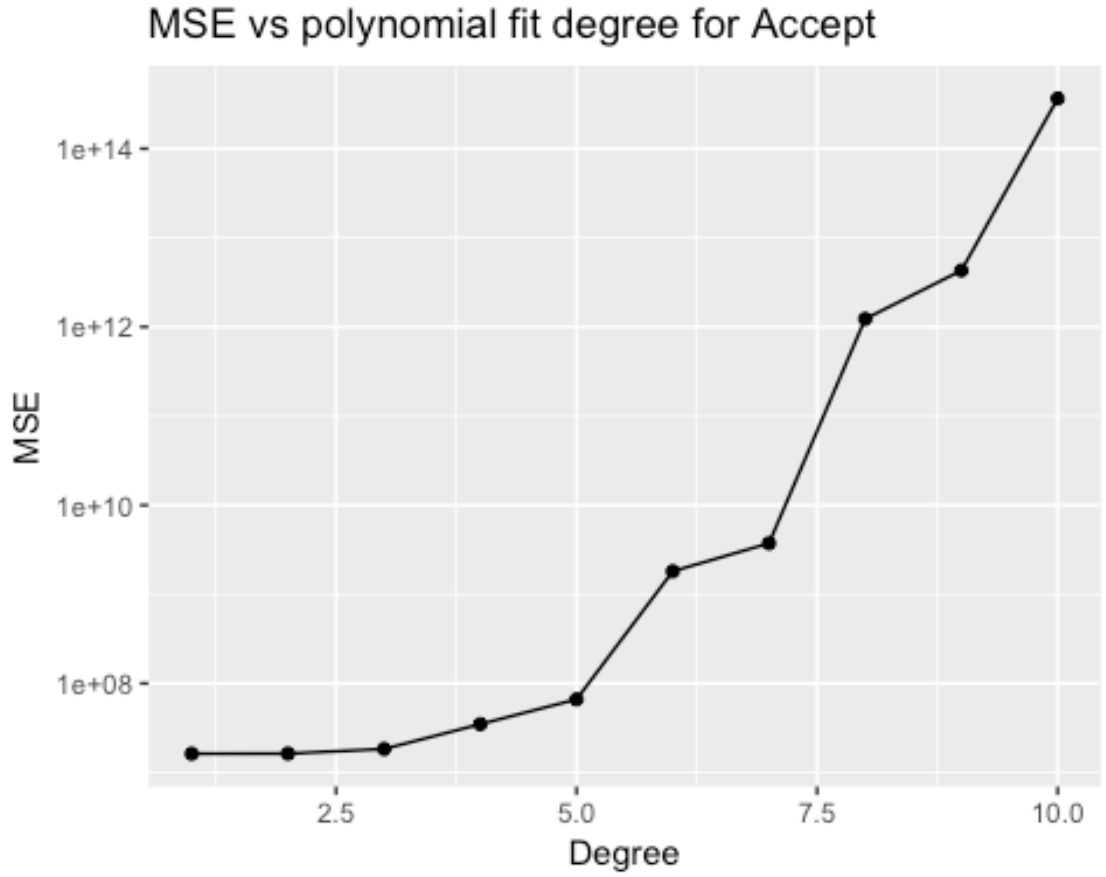


Accept

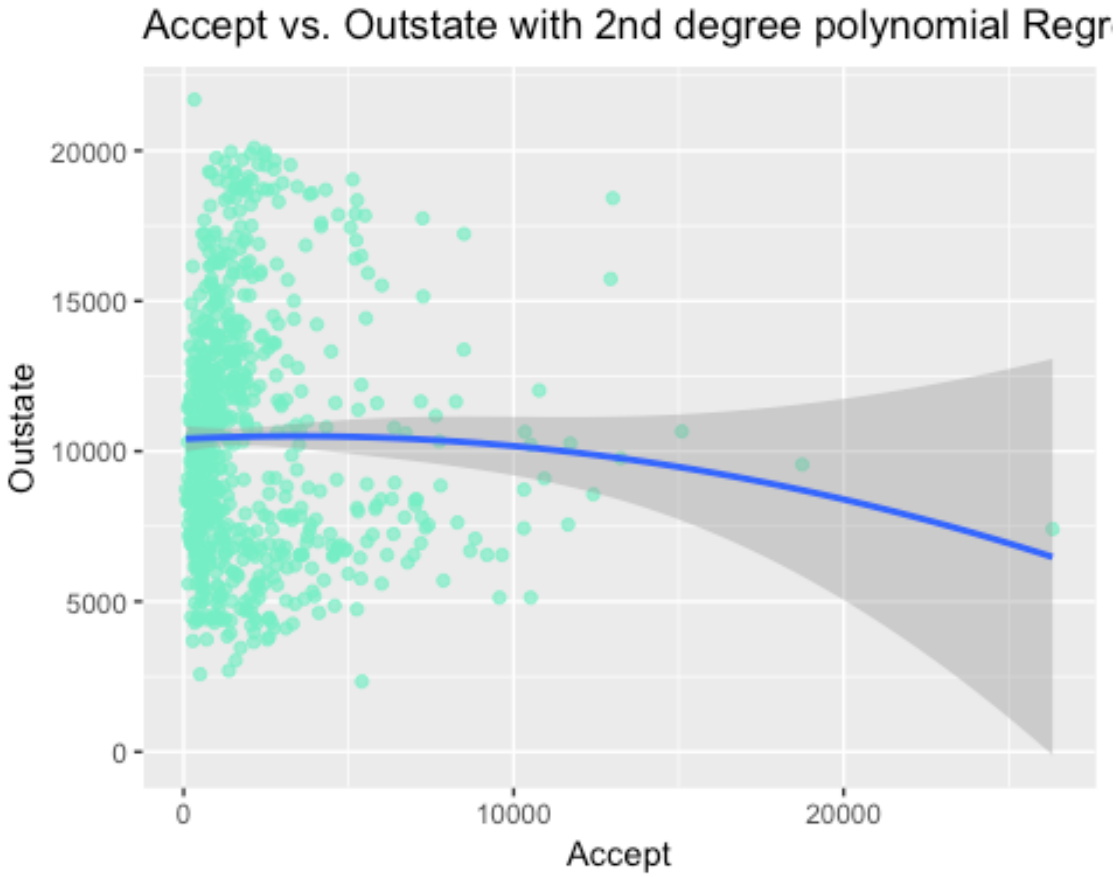
```
##
## Call:
## lm(formula = Outstate ~ Accept, data = c_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7957.7 -3080.4  -512.2  2425.3 11187.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.053e+04  1.871e+02  56.264   <2e-16 ***
## Accept      -4.227e-02  5.894e-02  -0.717    0.473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4024 on 775 degrees of freedom
## Multiple R-squared:  0.0006633,  Adjusted R-squared:  -0.0006262
## F-statistic: 0.5144 on 1 and 775 DF,  p-value: 0.4735
```


Linear model of Outstate regressed on Private
Residuals
10000
5000
0
-5000
9600
9900
10200
10500
Predicted Out of State Tuition

The relationship does not seem to be linear, yet a higher degree polynomials may have a better result.



From the plot, a second degree polynomial would be enough for linearity.



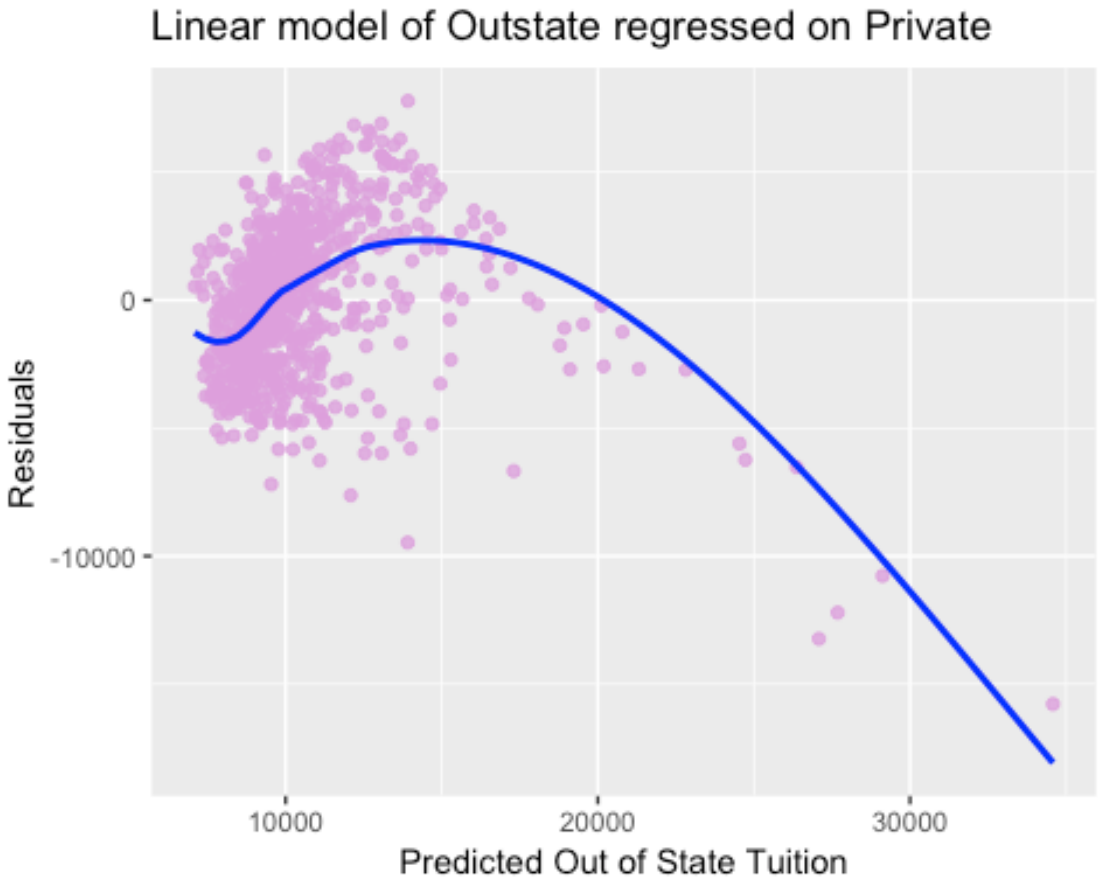
The relationship is now linear, yet has a negative direction, which means as number of acceptance increases, the out of state tuition will go down.

Expend

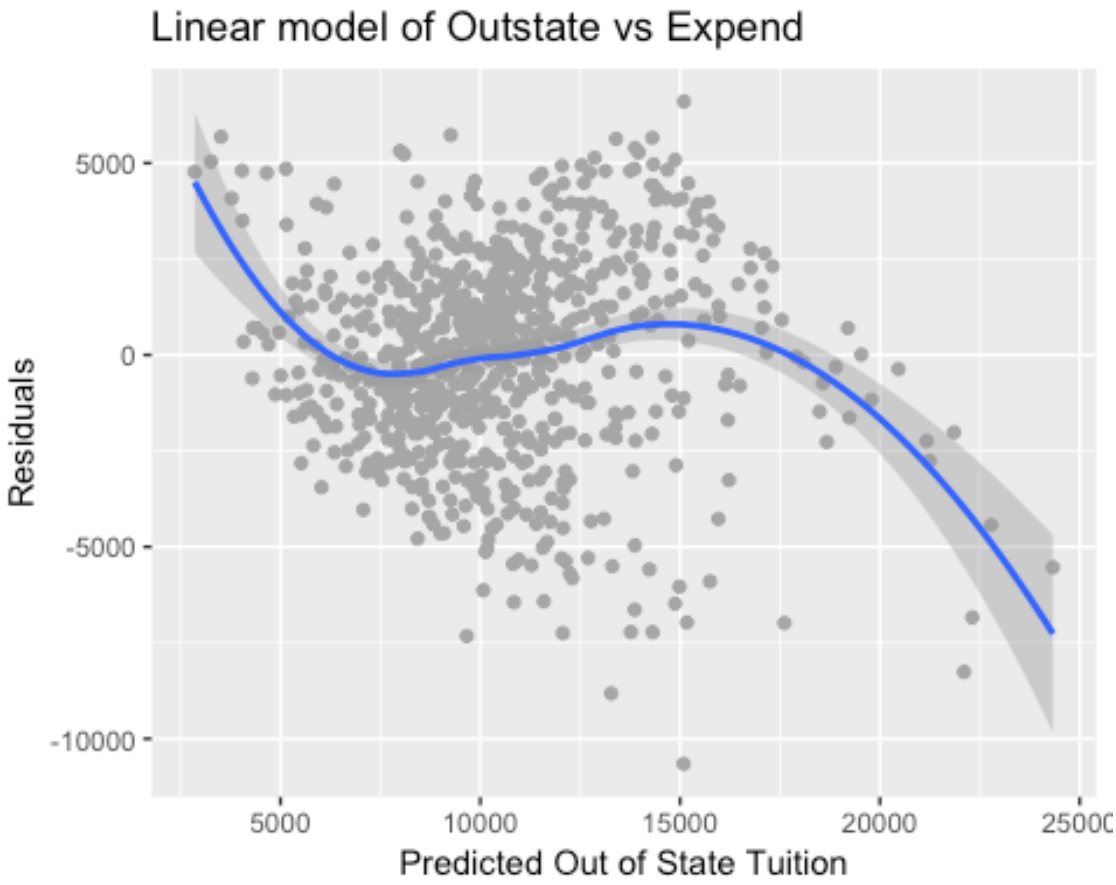
Finally, let's look at the expend variable.

```
##
## Call:
## lm(formula = Outstate ~ Expend, data = c_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15780.8  -2088.7     57.6   2010.8   7784.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.434e+03  2.248e+02   24.17   <2e-16 ***
## Expend      5.183e-01  2.047e-02   25.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2978 on 775 degrees of freedom
## Multiple R-squared:  0.4526, Adjusted R-squared:  0.4519
## F-statistic: 640.9 on 1 and 775 DF,  p-value: < 2.2e-16
```



From the plot, the model does not explain the relationship well. However, this may be transformed into log form.


Linear model of Outstate vs Expend
Residuals
5000
0
-5000
-10000
5000
10000
15000
20000
25000
Predicted Out of State Tuition

Now the relationship is reasonable: as instrucional expenditure per student increases, the out of state tuition rises.

Part 3: College (GAM) [3 points]

1. Split the data into a training set and a test set.

2. Estimate an OLS model on the training data, using out-of-state tuition (`Outstate`) as the response variable and the other six variables as the predictors. Interpret the results and explain your findings, using appropriate techniques (tables, graphs, statistical tests, etc.).

```
##
## Call:
## lm(formula = Outstate ~ Private + Room.Board + PhD + perc.alumni +
##     Expend + Grad.Rate, data = c_split$train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6654.5 -1273.7   156.7  1245.3  5663.5
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.499e+03  7.983e+02  -4.383 1.79e-05 ***
## PrivateYes   2.969e+03  3.555e+02   8.353 6.63e-15 ***
## Room.Board   8.438e-01  1.528e-01   5.522 9.12e-08 ***
## PhD          3.521e+01  9.815e+00   3.588 0.000408 ***
## perc.alumni  6.669e+01  1.271e+01   5.249 3.52e-07 ***
## Expend       2.027e-01  3.175e-02   6.385 9.57e-10 ***
## Grad.Rate    3.312e+01  9.281e+00   3.569 0.000437 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1948 on 227 degrees of freedom
## Multiple R-squared:  0.7726, Adjusted R-squared:  0.7665
## F-statistic: 128.5 on 6 and 227 DF,  p-value: < 2.2e-16
```

The R^2 is 0.7726, which means the model can explain 77.26% of the variance in the training data. All 6 predictors are very significant. In particular, private universities would charge more in tuition by 2969 dollars; 1 dollar increase in room-board costs would result in an increase of 0.8438 dollar of the out-of-state tuition. If percentage PhDs increase by 1 unit, tuition would rise by 29 dollar; likewise, if the percent of alumni donator increase by 1, the tuition would be 66.69 dollars more. 1 unit rise in instructional expenditure per student would cause the tuition to increase by 0.2027. Finally, 1 unit higher graduation rate would also increase the tuition by 33.12 dollars.

3. Estimate a GAM on the training data, using out-of-state tuition (`Outstate`) as the response variable and the other six variables as the predictors. You can select any non-linear method (or linear) presented in the readings or in-class to fit each variable. Plot the results, and explain your findings. Interpret the results and explain your findings, using appropriate techniques (tables, graphs, statistical tests, etc.).

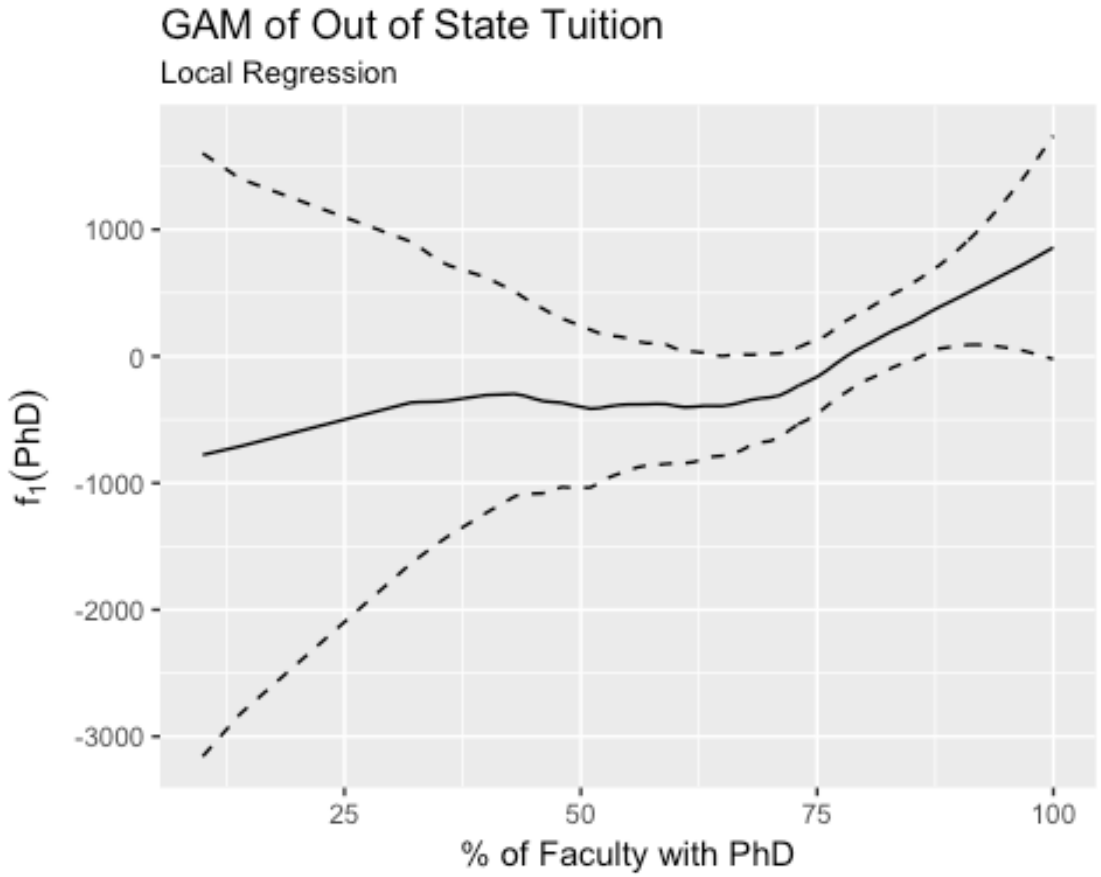
I am using GAM model to regress Outstate. With other variance have no transformations, I am using a 2nd degree polynomial of Room.Board and the log(Expend) instead.

```
##
## Call: gam(formula = Outstate ~ lo(PhD) + lo(perc.alumni) + log(Expend) +
##     lo(Grad.Rate) + Private + poly(Room.Board, 2), data = c_split$train)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -4627.2305 -1180.6469    -0.5081  1074.2161  5102.9368
##
## (Dispersion Parameter for gaussian family taken to be 3334280)
##
##     Null Deviance: 3787534384 on 233 degrees of freedom
## Residual Deviance: 726886404 on 218.004 degrees of freedom
## AIC: 4196.112
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                      Df     Sum Sq    Mean Sq F value    Pr(>F)
```

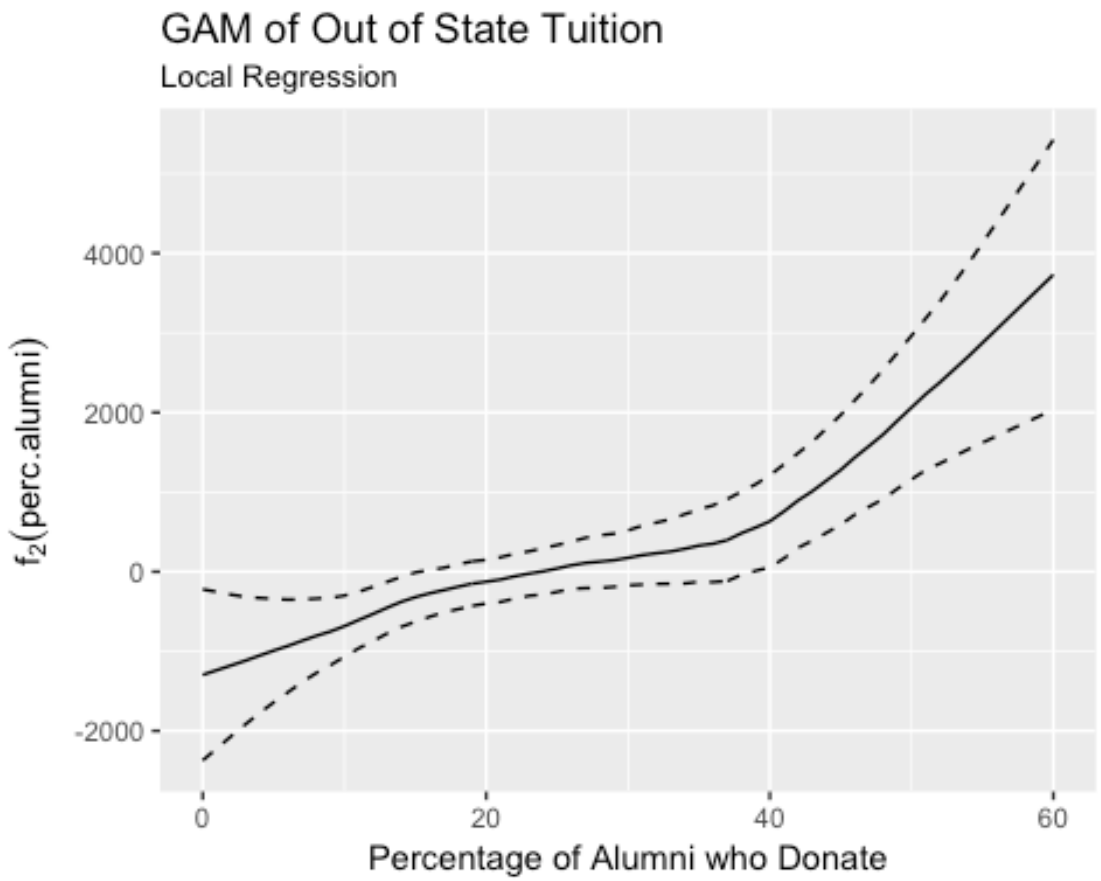
```
## lo(PhD)               1  622928435  622928435 186.826 < 2.2e-16 ***
## lo(perc.alumni)       1 1038633578 1038633578 311.502 < 2.2e-16 ***
## log(Expend)           1  761669446  761669446 228.436 < 2.2e-16 ***
## lo(Grad.Rate)         1  137125941  137125941  41.126 8.730e-10 ***
## Private               1  279167257  279167257  83.726 < 2.2e-16 ***
## poly(Room.Board, 2)   2   80455276   40227638  12.065 1.073e-05 ***
## Residuals           218  726886404    3334280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                     Npar Df Npar F  Pr(F)
## (Intercept)
## lo(PhD)                 2.8 1.0035 0.3892
## lo(perc.alumni)         2.5 2.7567 0.0537 .
## log(Expend)
## lo(Grad.Rate)           2.7 1.9405 0.1310
## Private
## poly(Room.Board, 2)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, unsurprisingly, all predictors are statistically significant at 0 level.

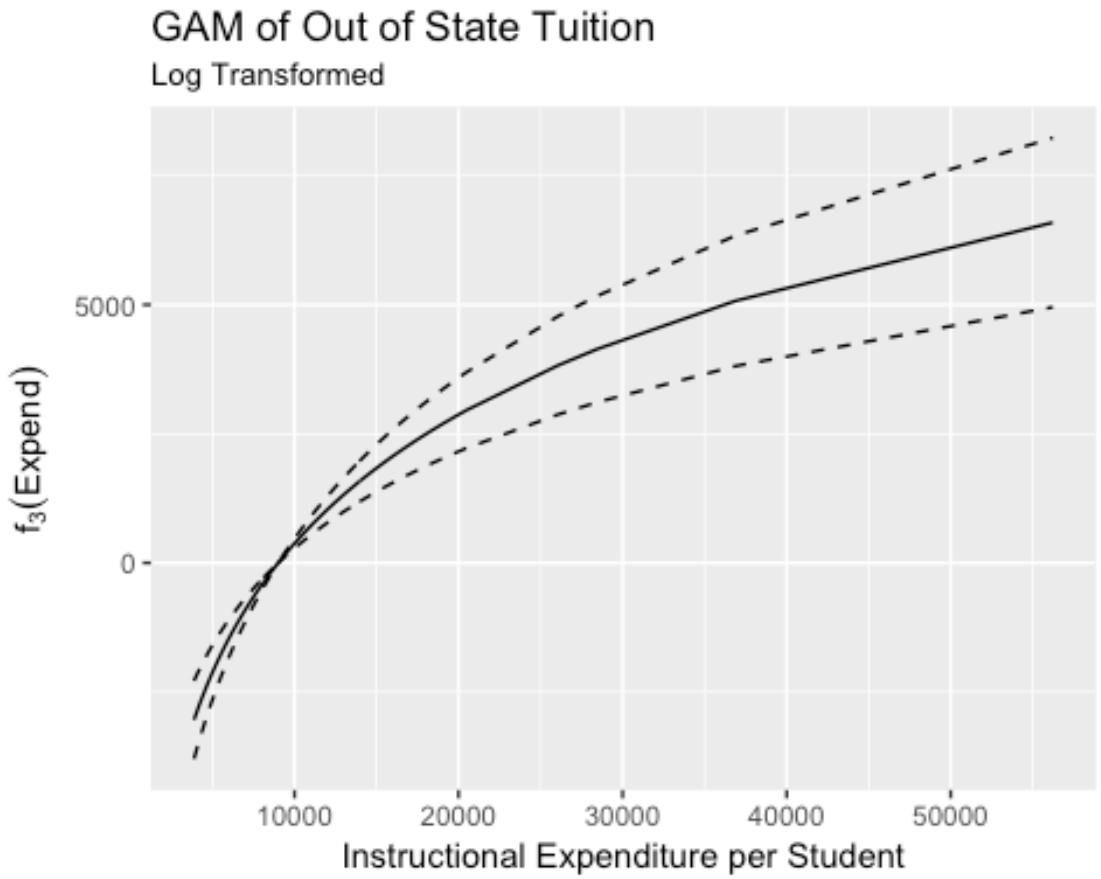
Now let's have a closer look at each response.



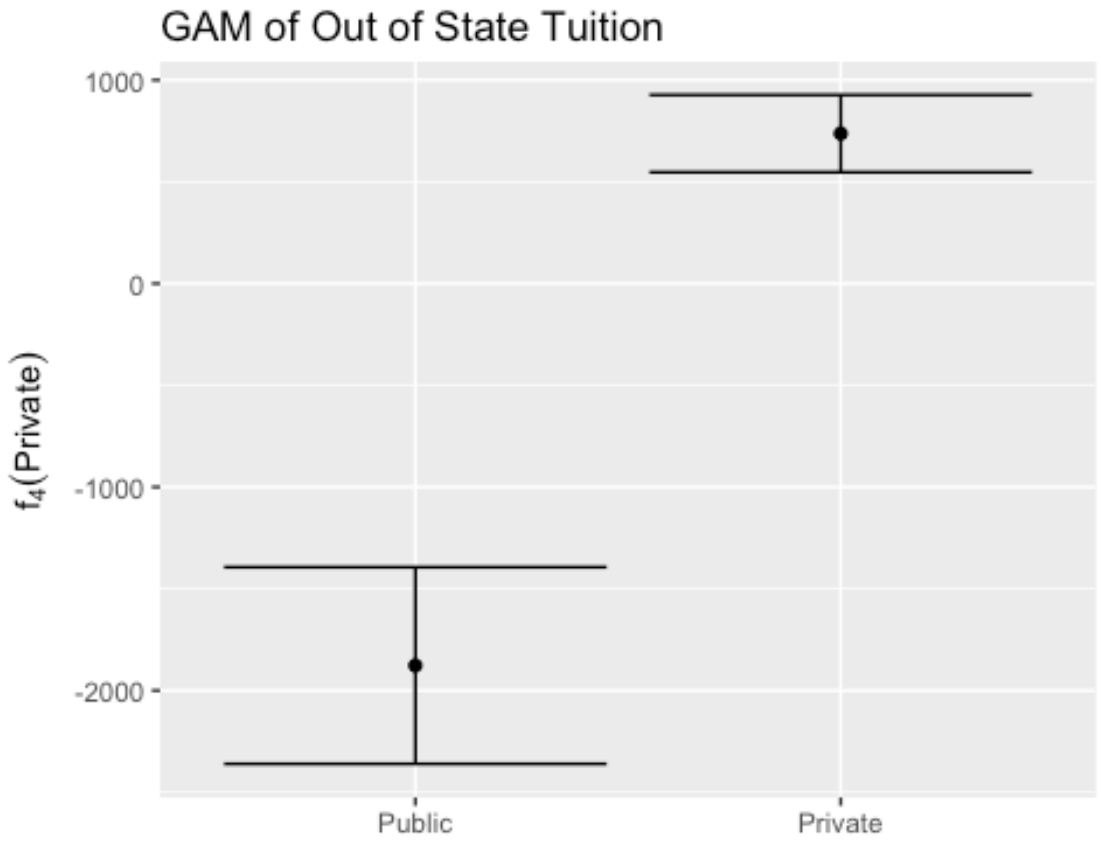
The overall trend tells us as percentage of PhD increases, tuition goes up.



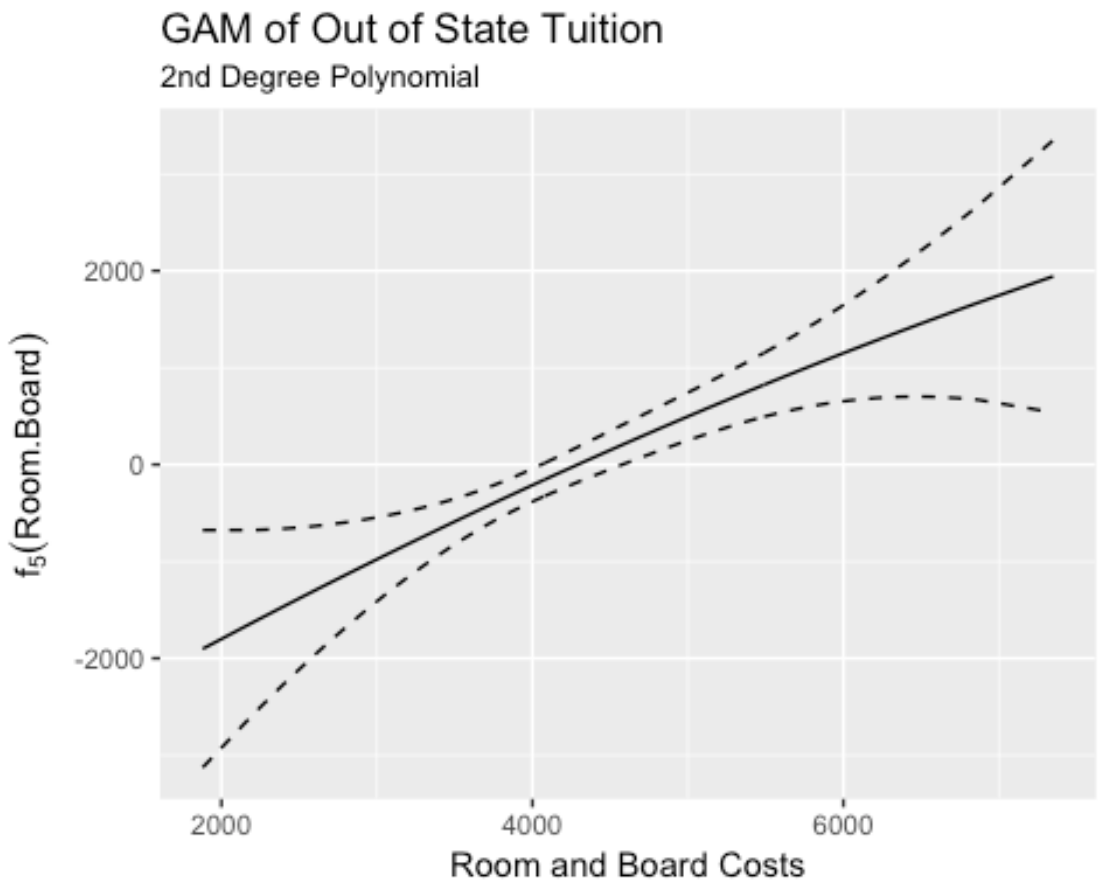
Similar upward trend for percentage of alumni, yet the confidence interval is smaller this time.



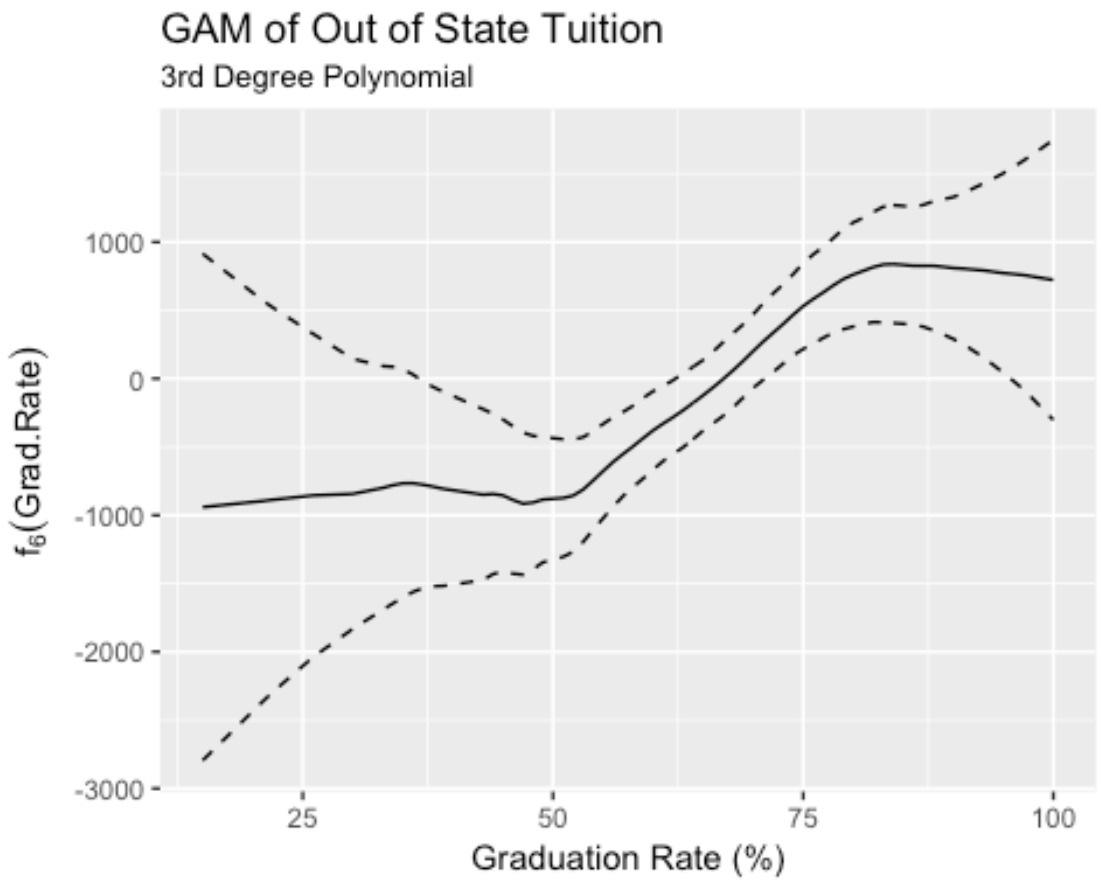
The expenditure per student also have a positive relationship with tuition. In particular, when instructional expeniture per student is low (below 15,000), the increasing rate is fast; as expenditure increases, the rate of increase in tuition slows down.



It is quite clear that being a public university has a negative effect on tuition fee while a private university has a positive effect.



There is a clear positive linear relaitonship between room and board costs and tuition.



There is no clear pattern in this trend. When graduation rate is low (below 50%), the effect is not trivial. After a small dip at 50% graduation rate, its increase has a positive relationship with tuition until it reaches around 87% graduation. Then the trend slightly goes down.

4. Use the test set to evaluate the model fit of the estimated OLS and GAM models, and explain the results obtained.

```
## [1] 4383889

## [1] 4105254
```

5. For which variables, if any, is there evidence of a non-linear relationship with the response?[2]

```
##
## Call: gam(formula = Outstate ~ lo(PhD) + lo(perc.alumni) + log(Expend) +
##     lo(Grad.Rate) + Private + poly(Room.Board, 2), data = c_split$train)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
```

[2] Hint: Review Ch. 7.8.3 from ISL on how you can use ANOVA tests to determine if a non-linear relationship is appropriate for a given variable.

```
## -4627.2305 -1180.6469    -0.5081  1074.2161  5102.9368
##
## (Dispersion Parameter for gaussian family taken to be 3334280)
##
##     Null Deviance: 3787534384 on 233 degrees of freedom
## Residual Deviance: 726886404 on 218.004 degrees of freedom
## AIC: 4196.112
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                       Df     Sum Sq    Mean Sq F value    Pr(>F)
## lo(PhD)                1  622928435  622928435 186.826 < 2.2e-16 ***
## lo(perc.alumni)        1 1038633578 1038633578 311.502 < 2.2e-16 ***
## log(Expend)            1  761669446  761669446 228.436 < 2.2e-16 ***
## lo(Grad.Rate)          1  137125941  137125941  41.126 8.730e-10 ***
## Private                1  279167257  279167257  83.726 < 2.2e-16 ***
## poly(Room.Board, 2)    2   80455276   40227638  12.065 1.073e-05 ***
## Residuals            218  726886404    3334280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                      Npar Df Npar F  Pr(F)
## (Intercept)
## lo(PhD)                  2.8 1.0035 0.3892
## lo(perc.alumni)          2.5 2.7567 0.0537 .
## log(Expend)
## lo(Grad.Rate)            2.7 1.9405 0.1310
## Private
## poly(Room.Board, 2)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the ANOVA for Nonparametric effects, no variable is statistically significant.