Bruce Campbell ST-617 Homework 2

Wed Jul 13 08:21:36 2016

Chapter 7

Problem 10

This question relates to the College data set.

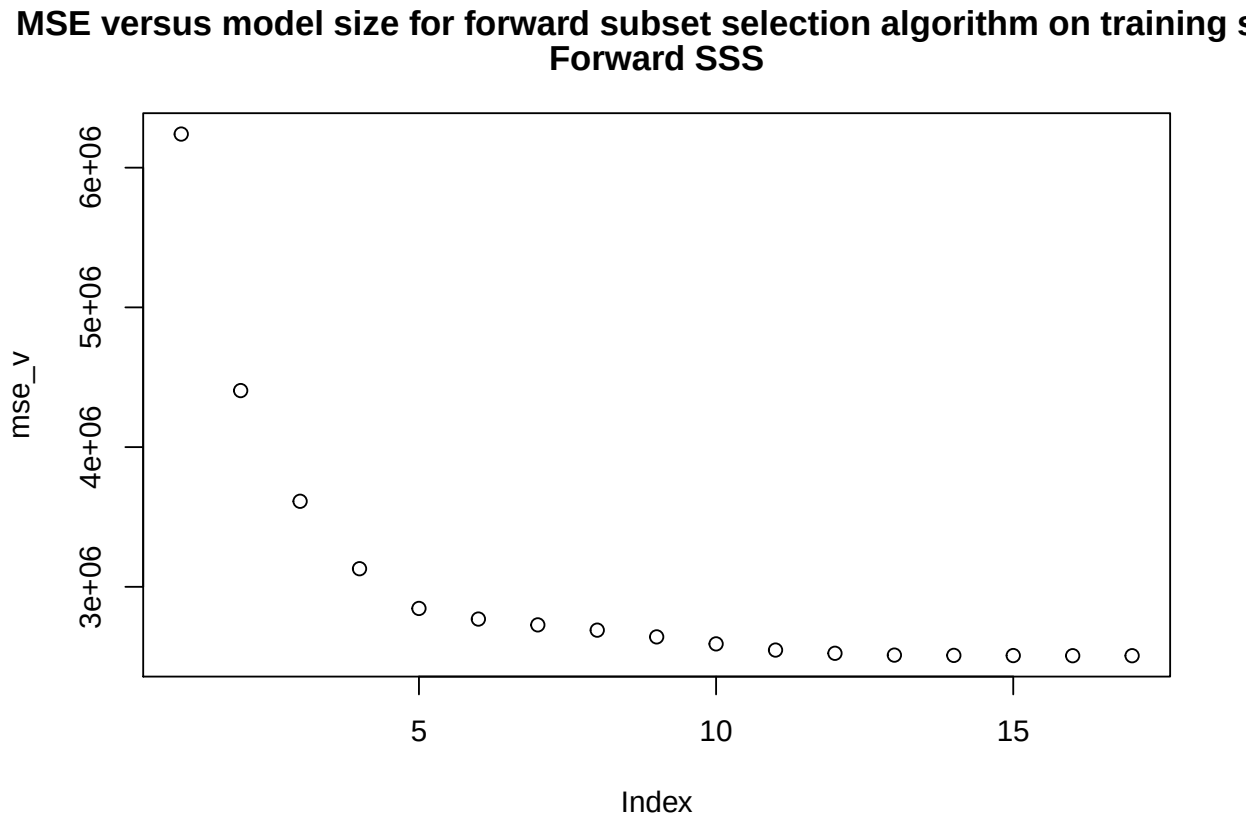
a)

Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

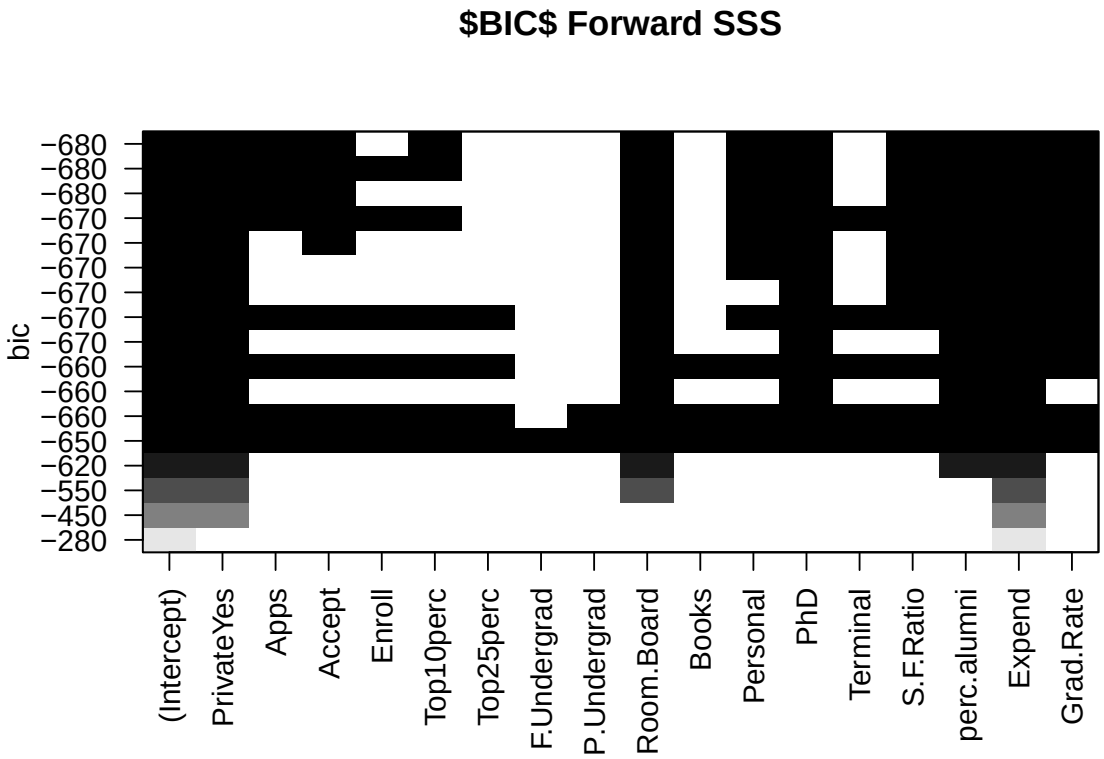
```
rm(list = ls())
library(ISLR)
attach(College)

train = sample(nrow(College), floor(nrow(College) * 2/3))
DF <- College
DFTrain <- DF[train, ]
DFTest <- DF[-train, ]

library(pander)
pander(names(DF))
```

Private, *Apps*, *Accept*, *Enroll*, *Top10perc*, *Top25perc*, *F.Undergrad*, *P.Undergrad*, *Outstate*, *Room.Board*, *Books*, *Personal*, *PhD*, *Terminal*, *S.F.Ratio*, *perc.alumni*, *Expend* and *Grad.Rate*

```
library(leaps)
regfit.full <- regsubsets(Outstate ~ ., data = DFTrain, method = "forward",
    nvmax = 18)

reg.summary <- summary(regfit.full)

mse_v <- reg.summary$rss/nrow(DF)

plot(mse_v)
title(c("MSE versus model size for forward subset selection algorithm on training set.",
    "Forward SSS"))
```




```
plot(regfit.full, scale = "bic")
title("$BIC$ Forward SSS")
```



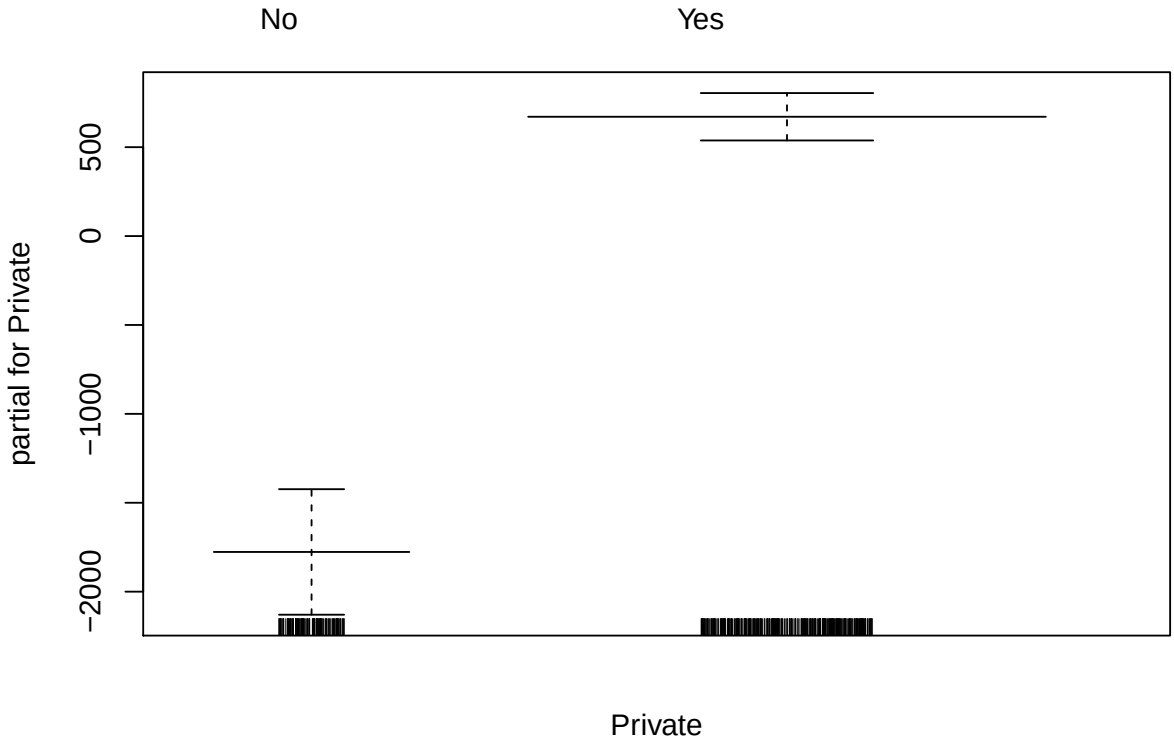

```
model_fss8 <- coef(regfit.full, 8)
pander(names(model_fss8))
```

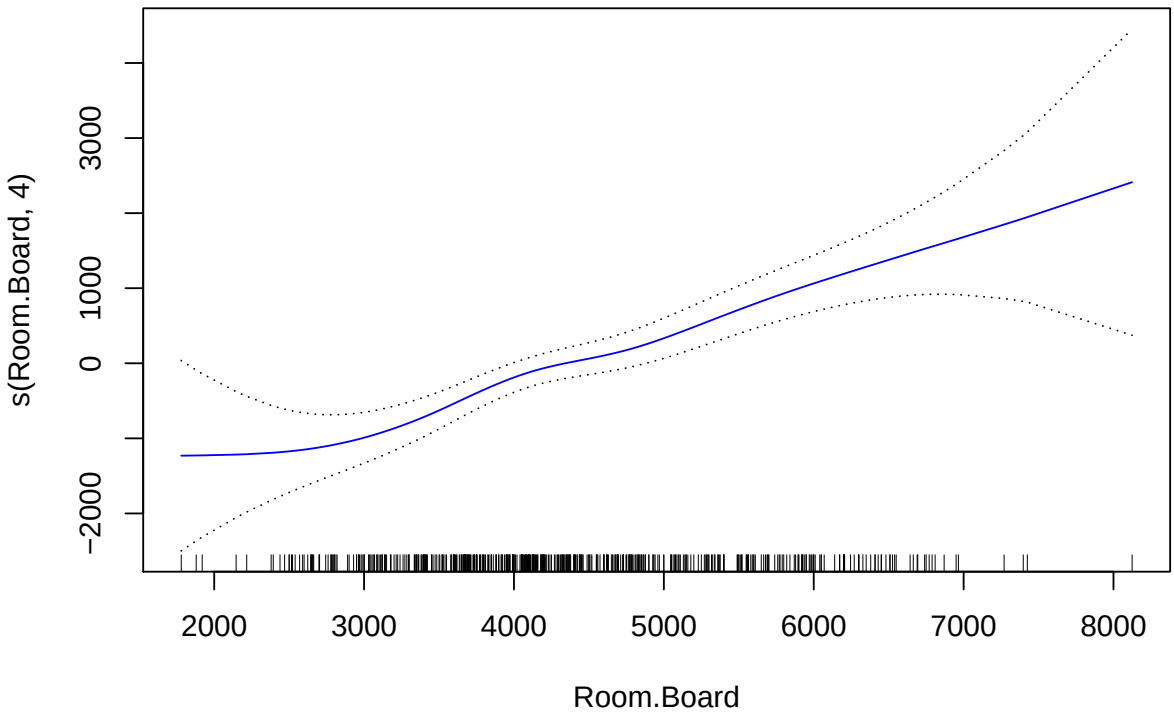
(Intercept), *PrivateYes*, *Room.Board*, *Personal*, *PhD*, *S.F.Ratio*, *perc.alumni*, *Expend* and *Grad.Rate*

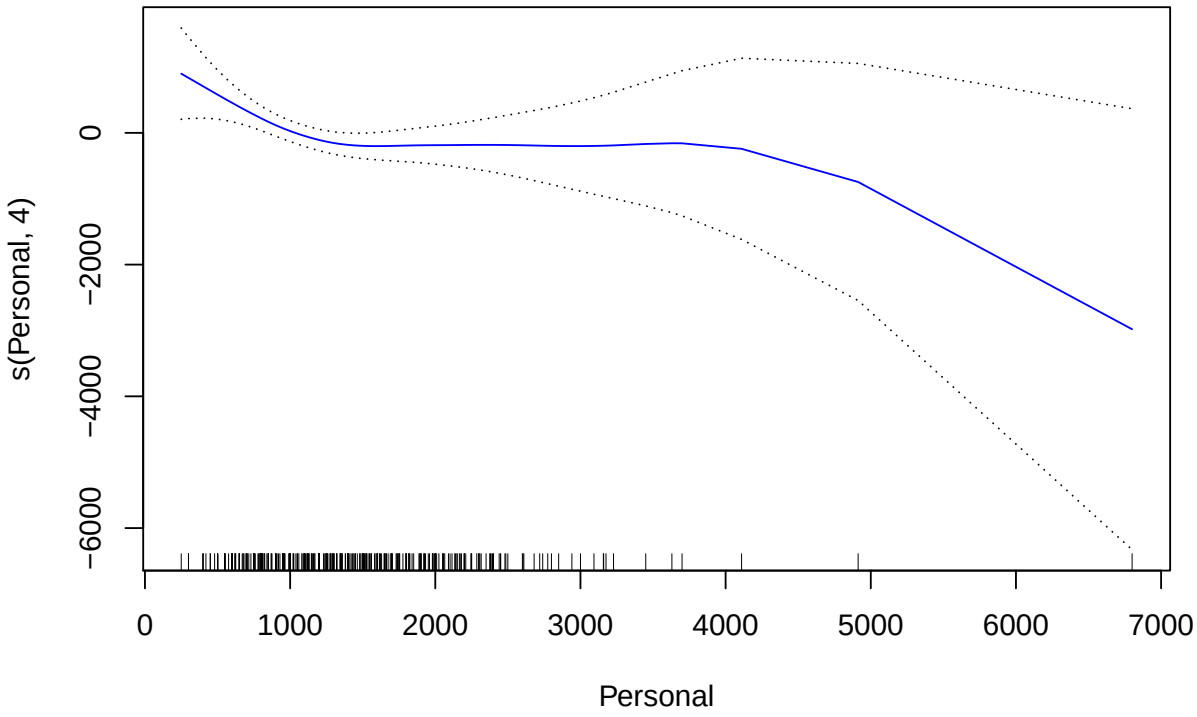
b)

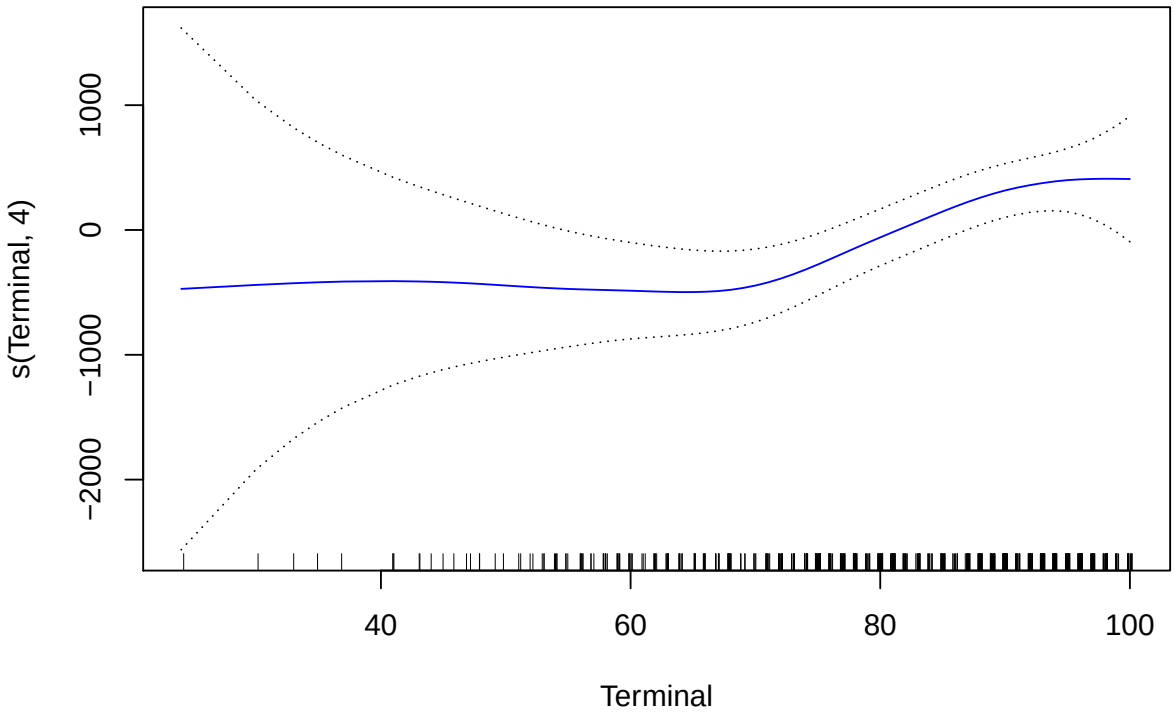
Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

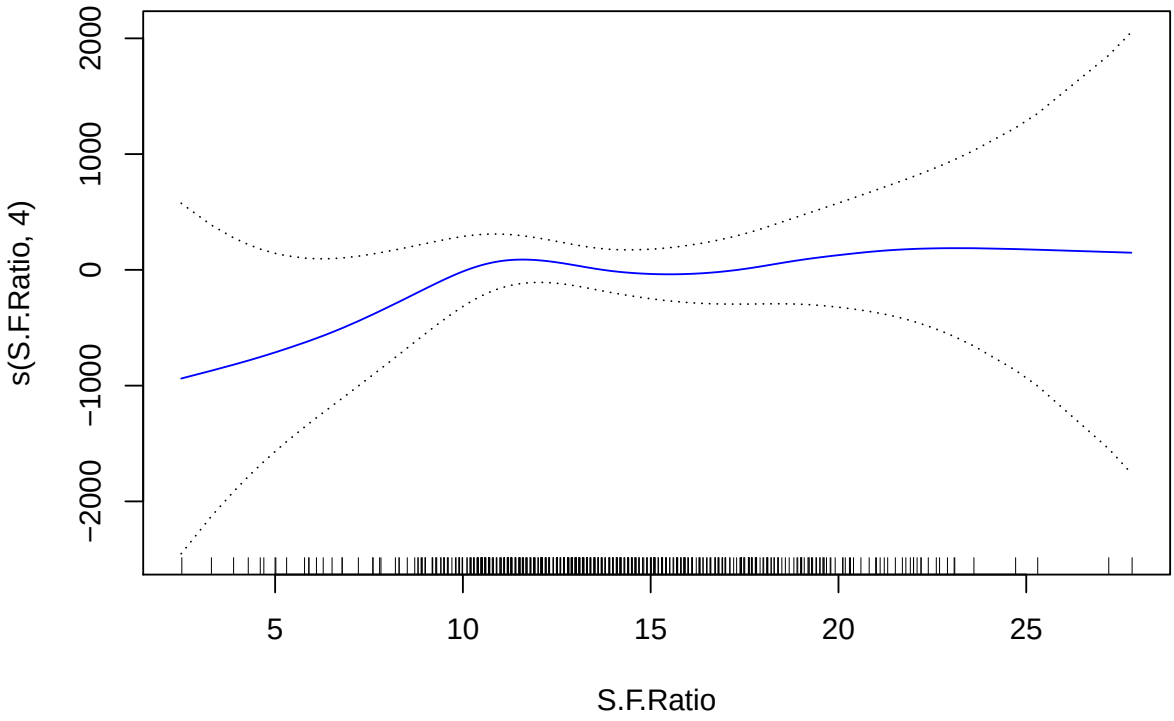
```
library(gam)

gam.fit = gam(Outstate ~ Private + s(Room.Board, 4) + s(Personal, 4) + s(Terminal,
    4) + s(S.F.Ratio, 4) + s(perc.alumni) + s(Expend, 4) + s(Grad.Rate, 4),
    data = DFTrain)

plot(gam.fit, se = TRUE, col = "blue ")
```

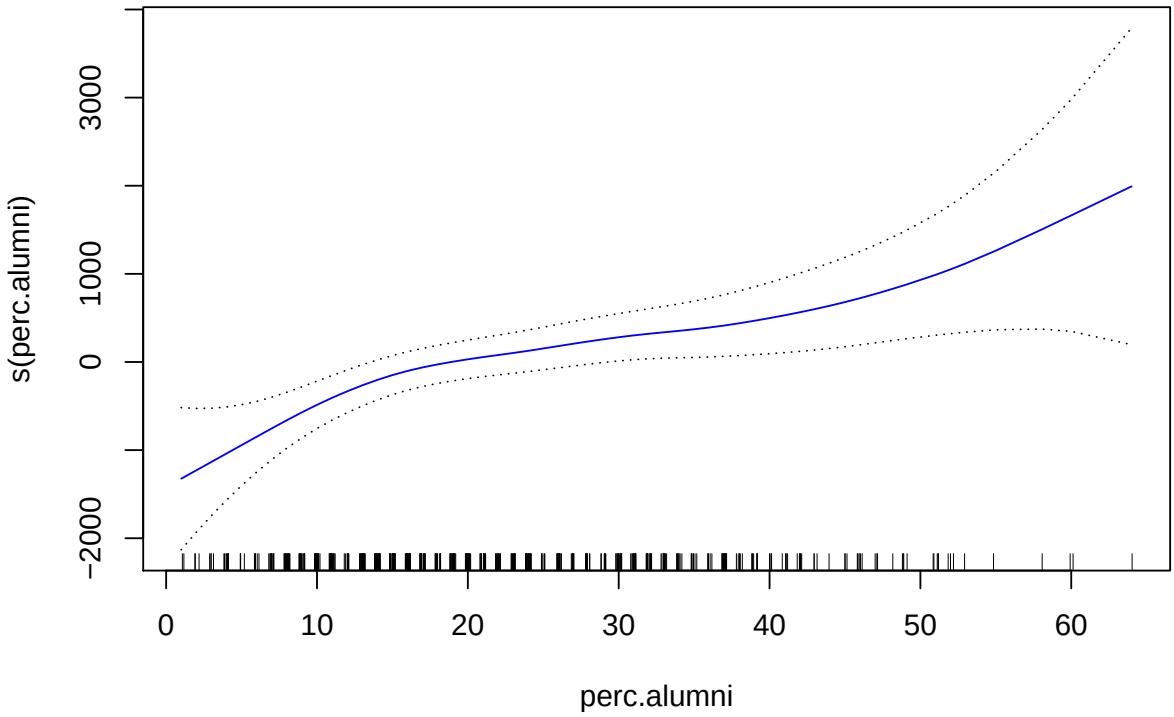

No
Yes
partial for Private
500
0
−1000
−2000
Private

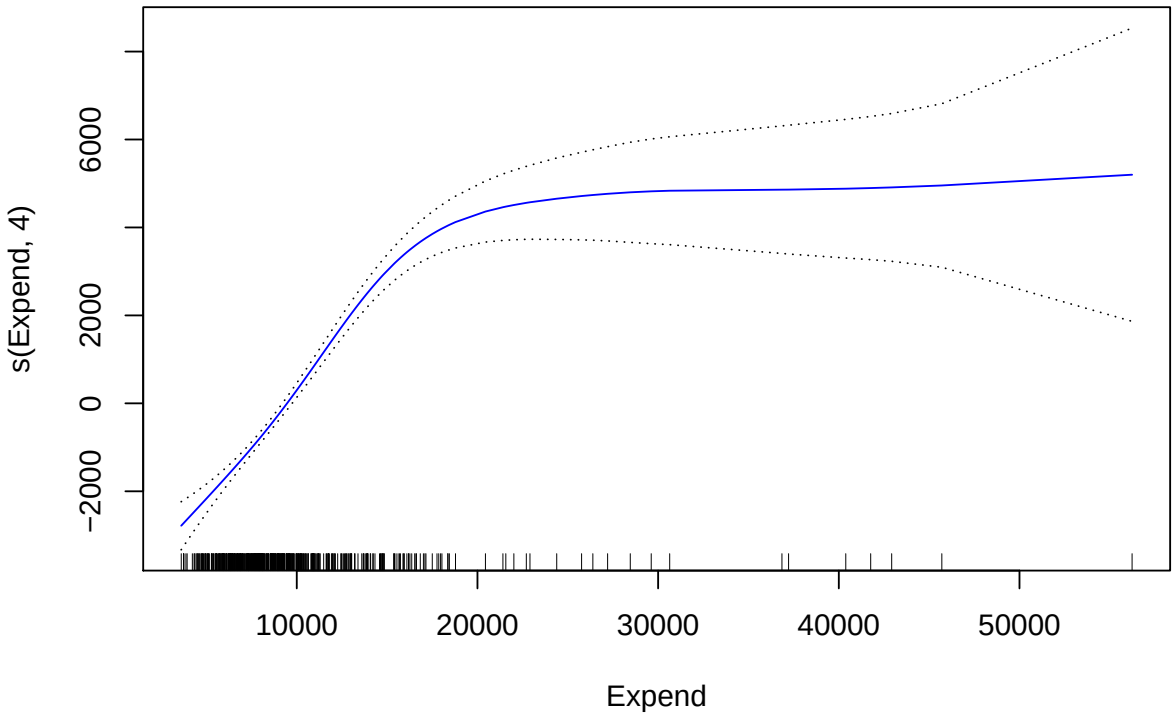

s(Room.Board, 4)
3000
1000
0
−2000
2000
3000
4000
5000
6000
7000
8000
Room.Board

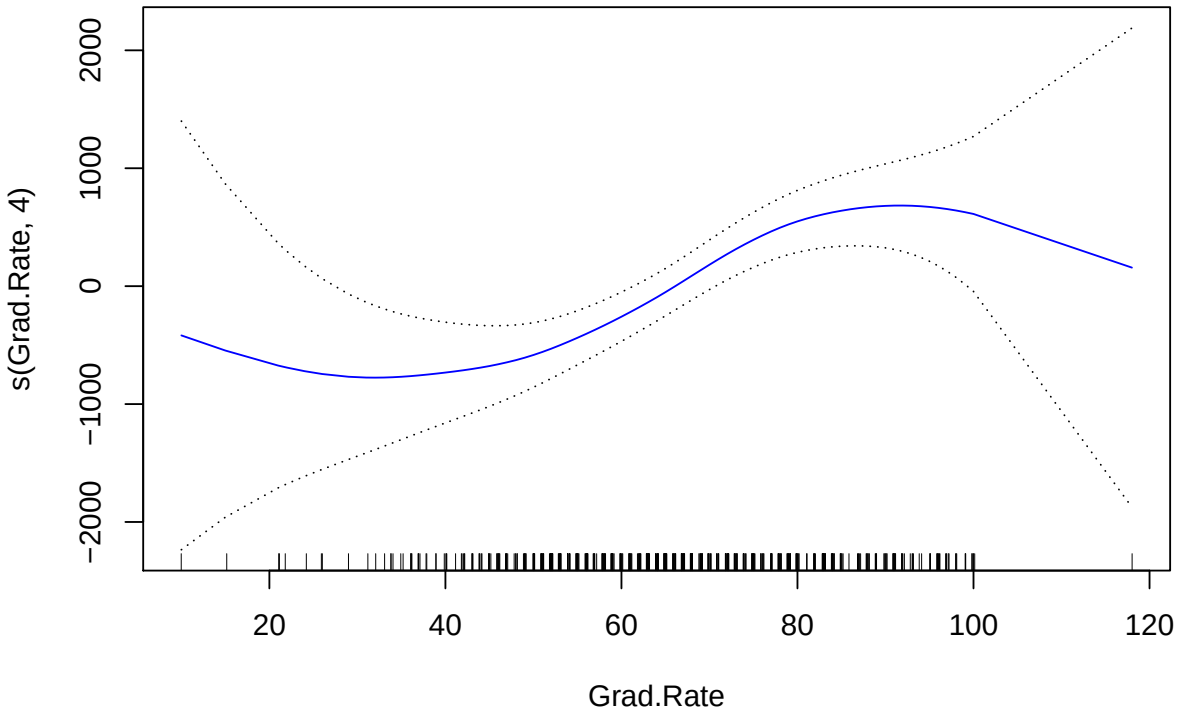

s(Personal, 4)
0
−2000
−4000
−6000
0
1000
2000
3000
4000
5000
6000
7000
Personal


s(Terminal, 4)
1000
0
−1000
−2000
40
60
80
100
Terminal


s(S.F.Ratio, 4)
2000
1000
0
−1000
−2000
5
10
15
20
25
S.F.Ratio


s(perc.alumni)
3000
1000
0
−2000
0
10
20
30
40
50
60
perc.alumni


s(Expend, 4)
−2000
0
2000
6000
10000
20000
30000
40000
50000
Expend

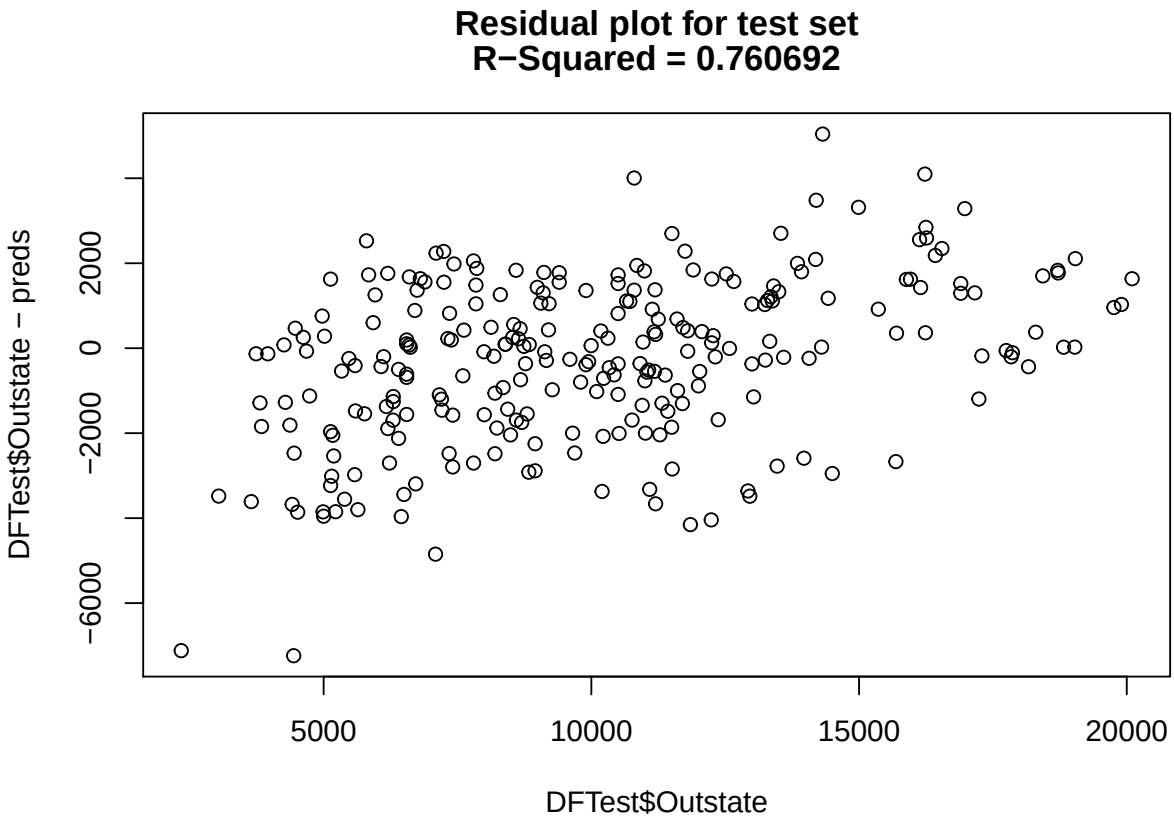


We have used the gam function to fit a univariate smmothing spline with 4 degrees of freedon to each perdictor. The plots show the univariate fits. One of the predictors selected by the forward SSS algorithm is a factor and is not fit to a smoothing spline. There is strong evidence of non linear relationships in the data. The variables $Personal, S.F.Ratio, perc.alumni, expend$ show this particularly.

c)

Evaluate the model obtained on the test set, and explain the results obtained.

```
preds = predict(gam.fit, newdata = DFTest)

RSS <- sum((preds - DFTest$Outstate)^2)
TSS <- sum((DFTest$Outstate - mean(DFTest$Outstate))^2)
RS2_Test <- 1 - (RSS/TSS)

plot(DFTest$Outstate, DFTest$Outstate - preds)
title(c("Residual plot for test set", sprintf("R-Squared = %f", RS2_Test)))
```




```
preds = predict(gam.fit, newdata = DFTrain)
RSS <- sum((preds - DFTrain$Outstate)^2)
TSS <- sum((DFTrain$Outstate - mean(DFTrain$Outstate))^2)
RS2_Train <- 1 - (RSS/TSS)
```

We see no significant trend in the residual plot, indicating that there is no unaccounted for non-linear relationships in the model. We also see that the training and test set R^2 statistic indicate a resonable fit. As expected the test R^2 statistic is below the training R^2 value.

d)

For which variables, if any, is there evidence of a non-linear relationship with the response?

The variables $Personal, S.F.Ratio, perc.alumni, expend$ particularly show a non -linear relationship with the response.