p8106_hw2_dd2948

David DeStephano

March 15, 2020

```
library(caret)
library(splines)
library(lasso2)
library(mgcv)
library(tidyverse)
library(ggplot2)
library(earth)
library(pdp)
library(tinytex)
```

Import the data

```
college<-read_csv("College.csv")
```

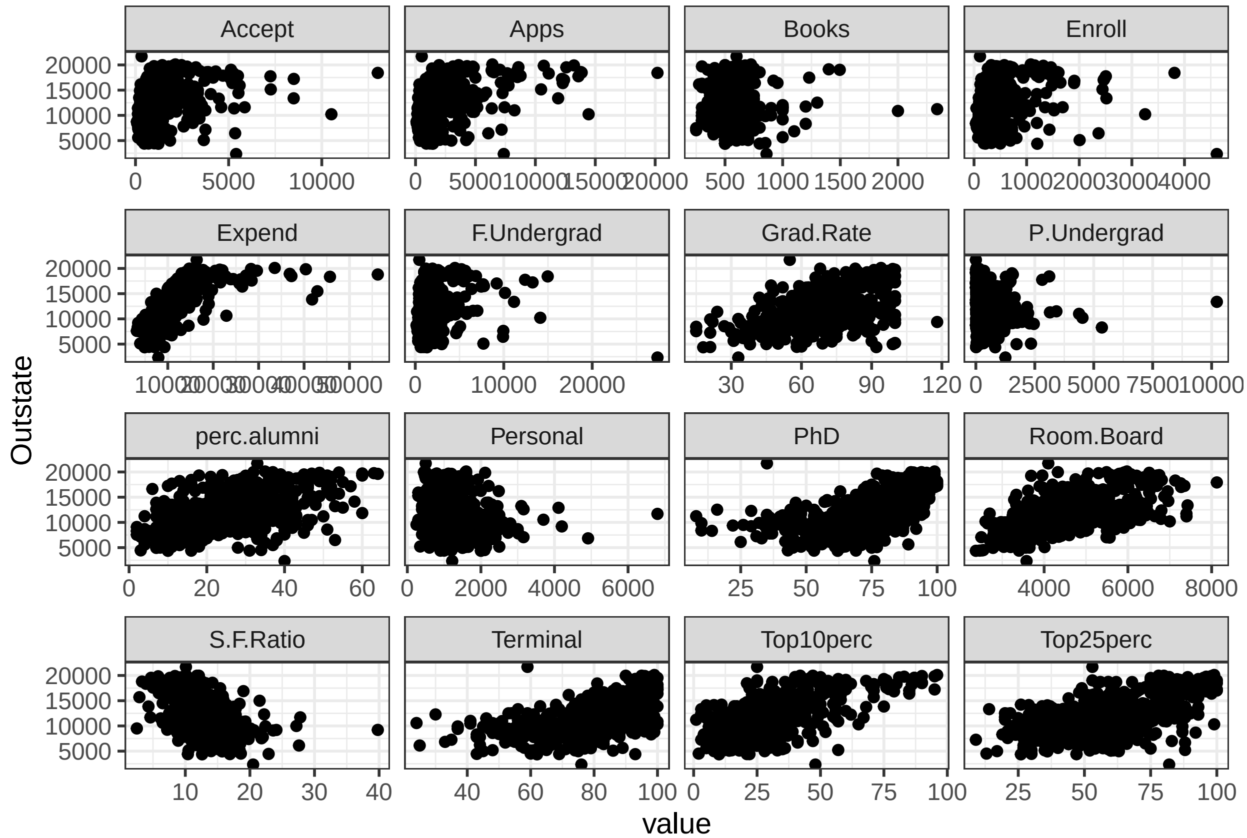
Set seed to 2

```
set.seed(2)
```

Scatter plots of response vs. predictors.

```
college %>%
  gather(-Outstate, -College, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = Outstate)) +
    geom_point() +
    facet_wrap(~ var, scales = "free_x") +
    theme_bw()
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



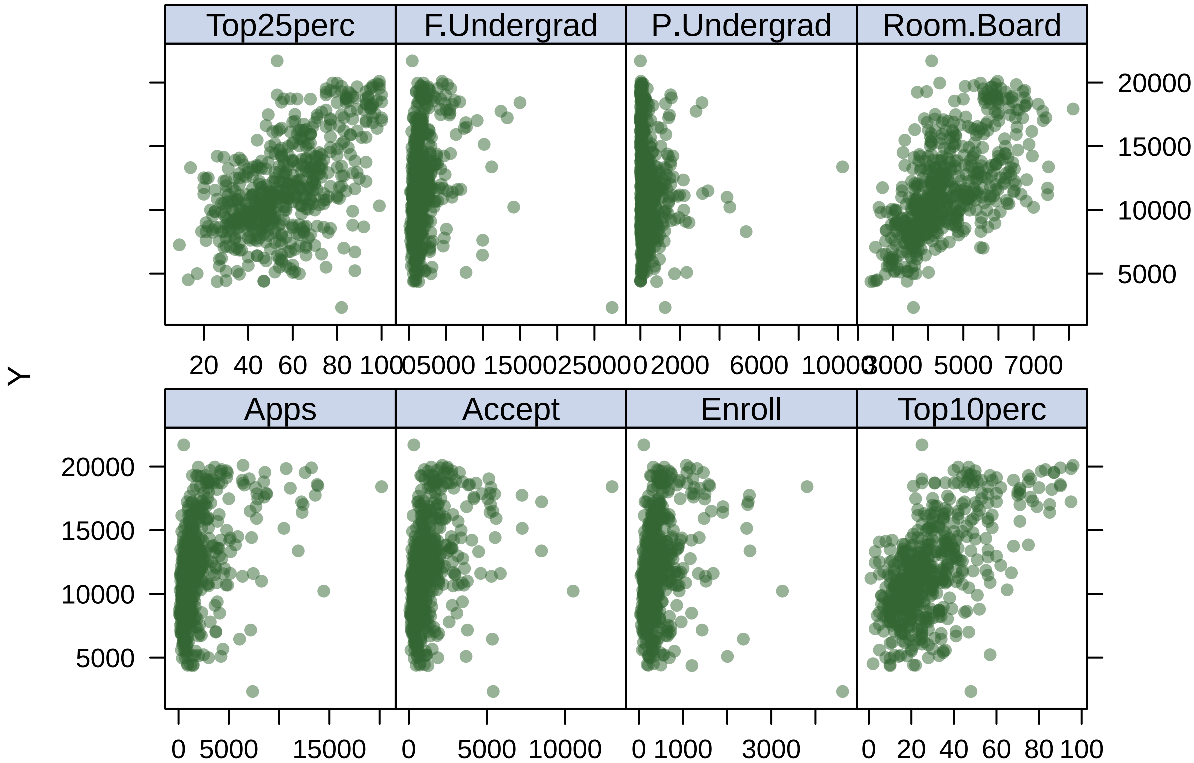
Alternative plot

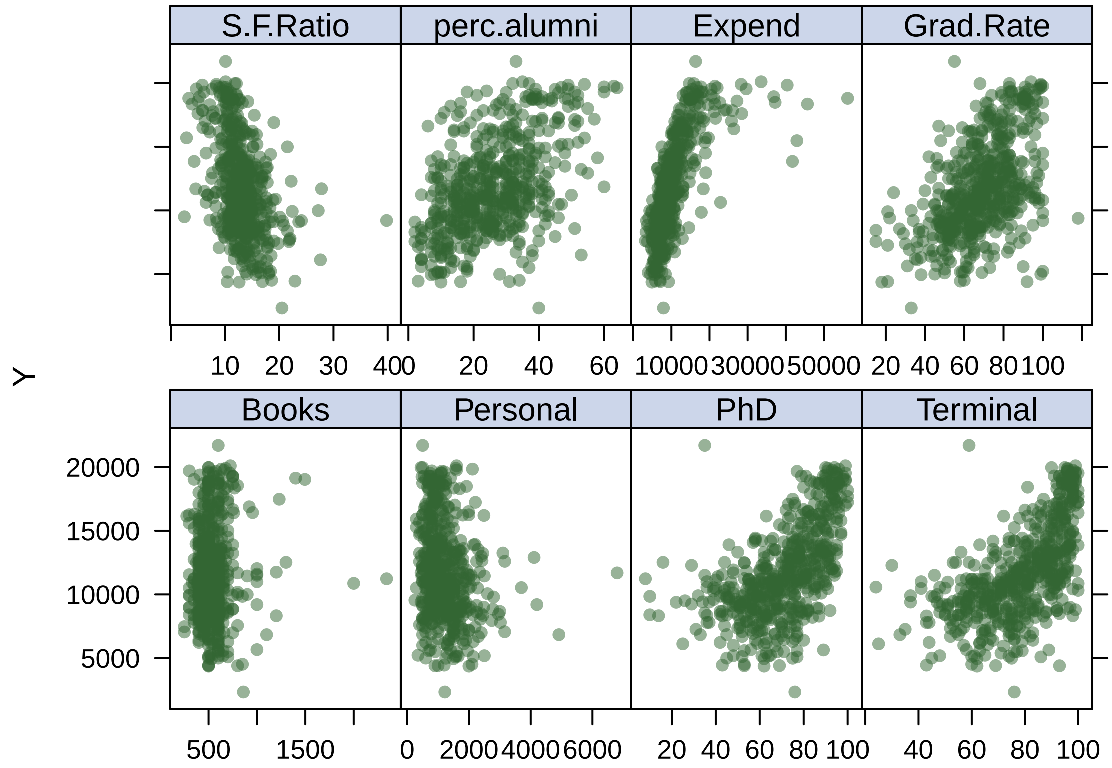
```r
map(college, ~sum(is.na(.)))
```

```
## $College
## [1] 0
##
## $Apps
## [1] 0
##
## $Accept
## [1] 0
##
## $Enroll
## [1] 0
##
## $Top10perc
## [1] 0
##
## $Top25perc
## [1] 0
##
## $F.Undergrad
```

```
## [1] 0
##
## $P.Undergrad
## [1] 0
##
## $Outstate
## [1] 1
##
## $Room.Board
## [1] 0
##
## $Books
## [1] 0
##
## $Personal
## [1] 0
##
## $PhD
## [1] 0
##
## $Terminal
## [1] 0
##
## $S.F.Ratio
## [1] 0
##
## $perc.alumni
## [1] 0
##
## $Expend
## [1] 0
##
## $Grad.Rate
## [1] 0
```

```r
college2<-college %>% select(-College) %>% select(-Outstate,Outstate) %>% filter(!is.na(Outstate))


# matrix of predictors
x <-model.matrix(Outstate~.,college2)[,-1]
# vector of response
y <- college2$Outstate


theme1 <-trellis.par.get()
theme1$plot.symbol$col <-rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <-rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <-rgb(.0, .2, .6, .2)
trellis.par.set(theme1)

featurePlot(x, y, plot = "scatter", labels =c("","Y"),type =c("p"), layout =c(4, 2))
```


Top25perc
F.Undergrad
P.Undergrad
Room.Board
Apps
Accept
Enroll
Top10perc
Y



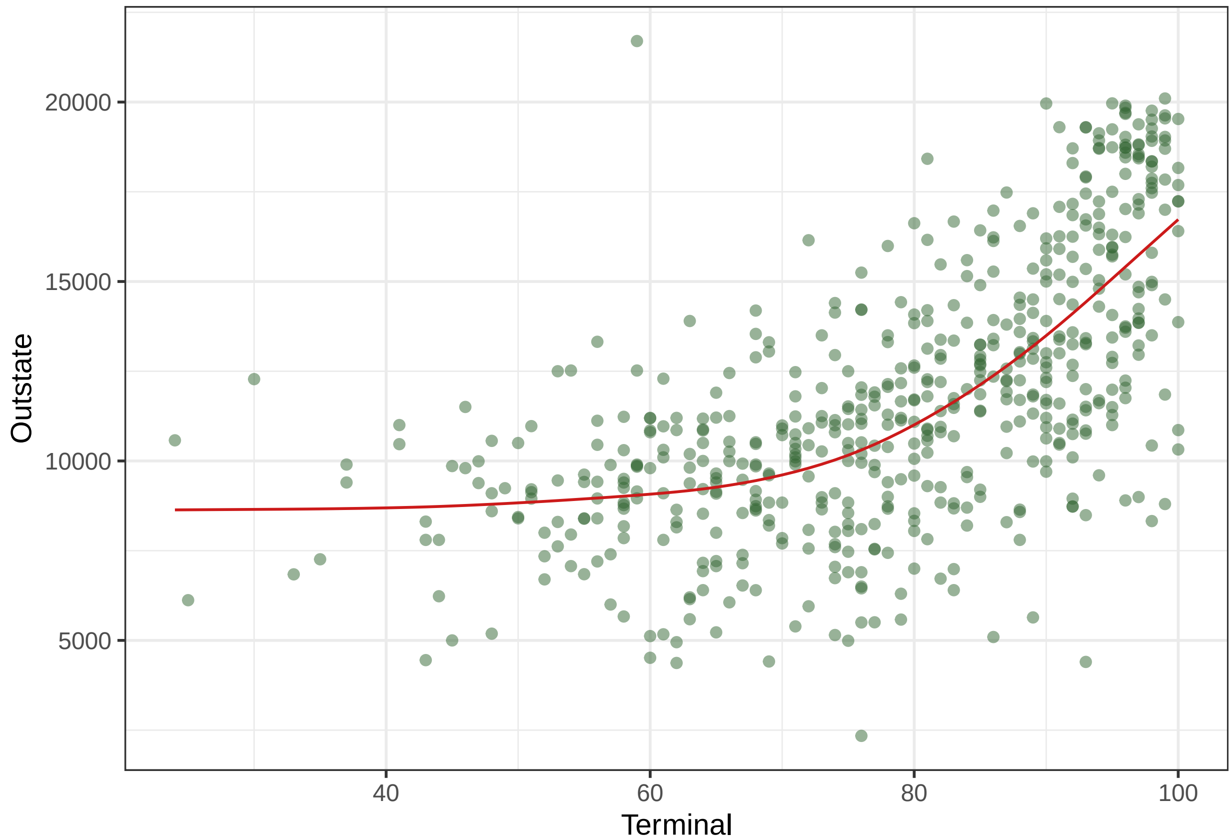
b) Fit a smoothing spline model using Terminal as the only predictor of Outstate for a range of degrees of freedom, as well as the degree of freedom obtained by generalized cross-validation, and plot the resulting fits. Describe the results obtained.

```
fit.ss <-smooth.spline(college2$Terminal, college2$Outstate, cv=FALSE) #cv=FALSE gives us Generalized c
fit.ss$df
```

```
## [1] 4.468629
```

```
Terminallims <-range(college2$Terminal)
Terminal.grid <-seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <-predict(fit.ss,
                  x = Terminal.grid)

pred.ss.df <-data.frame(pred = pred.ss$y,
                        Terminal = Terminal.grid)

p <-ggplot(data = college2,aes(x = Terminal, y =Outstate)) +
  geom_point(color =rgb(.2, .4, .2, .5))
```

```r
p+geom_line(aes(x = Terminal, y = pred), data = pred.ss.df,color =rgb(.8, .1, .1, 1))+ theme_bw()
```

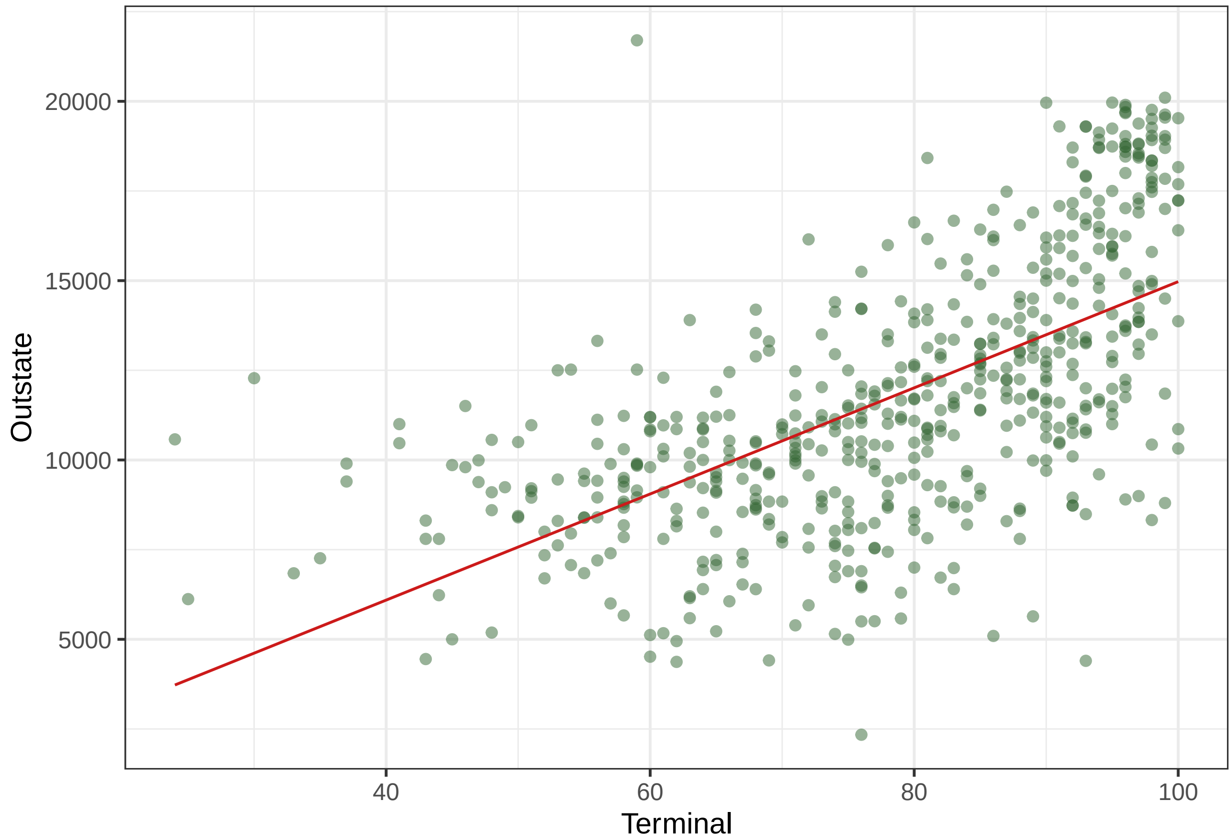


Fit a spline using a range of degrees of freedom

```r
#1 degree of freedom

fit.ss <-smooth.spline(college2$Terminal, college2$Outstate, df=2) #cv=FALSE gives us Generalized cross
fit.ss$df
```

```
## [1] 2.000314
```

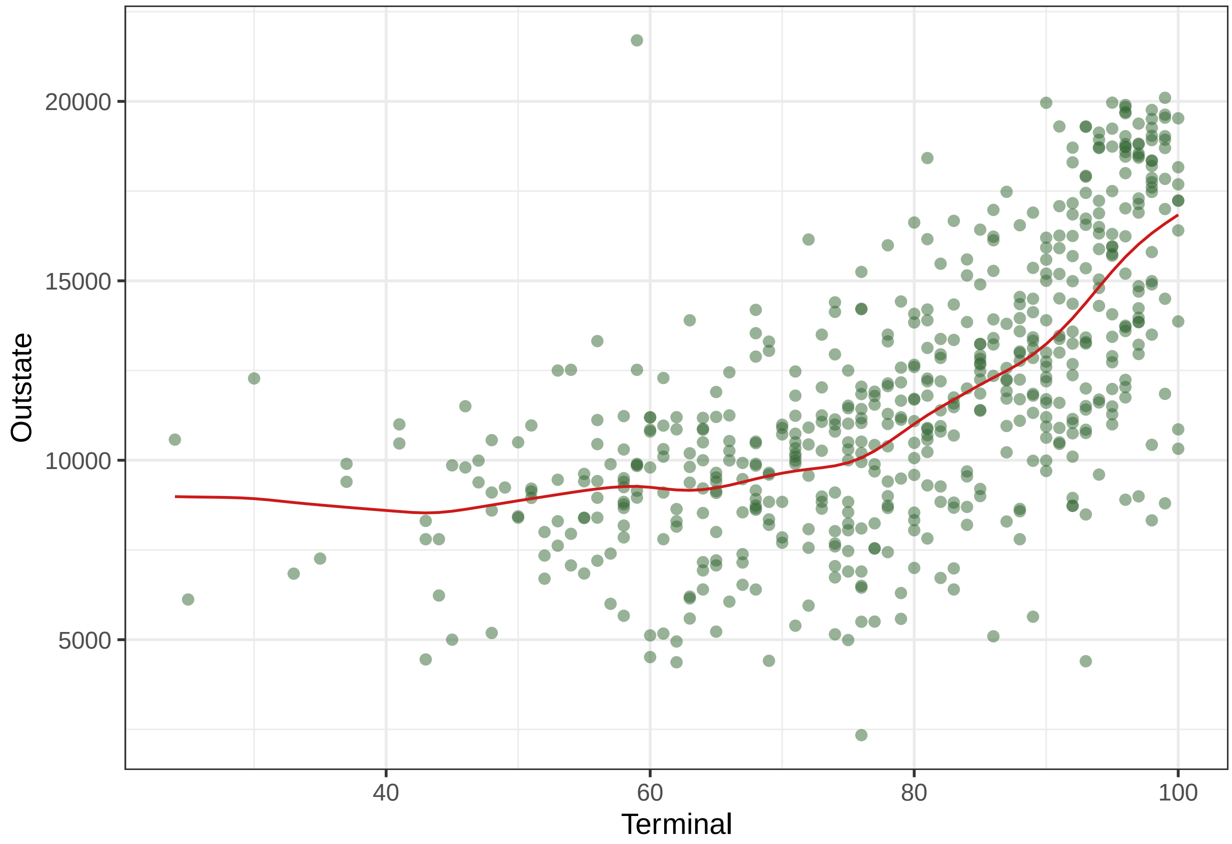
```r
pred.ss <-predict(fit.ss,
                  x = Terminal.grid)
pred.ss.df <-data.frame(pred = pred.ss$y,
                        Terminal = Terminal.grid)
p <-ggplot(data = college2,aes(x = Terminal, y =Outstate)) +
  geom_point(color =rgb(.2, .4, .2, .5))
p+geom_line(aes(x = Terminal, y = pred), data = pred.ss.df,color =rgb(.8, .1, .1, 1))+ theme_bw()
```




```r
#10 degrees of freedom
fit.ss <-smooth.spline(college2$Terminal, college2$Outstate, df=10)
fit.ss$df
```

```
## [1] 10.00143
```

```r
pred.ss <-predict(fit.ss,
                  x = Terminal.grid)
pred.ss.df <-data.frame(pred = pred.ss$y,
                        Terminal = Terminal.grid)
p <-ggplot(data = college2,aes(x = Terminal, y =Outstate)) +
  geom_point(color =rgb(.2, .4, .2, .5))
p+geom_line(aes(x = Terminal, y = pred), data = pred.ss.df,color =rgb(.8, .1, .1, 1))+ theme_bw()
```



The smoothing spline was fit using the smooth.spline function. When using generalized cross validation, the resulting degrees of frredom is 4.47. When picking your own degrees of freedom, larger values make the line much more wiggly, while lower degrees of freedom are more linear. It is important to model terminal with the cross validated degrees of freedom, as there appears to be no real affect of the percent of staff with a terminal degreee until the 70th percentile, so a linear term would not be very accurate.

(c) Fit a generalized additive model (GAM) using all the predictors. Plot the results and explain your findings.

```
gam.m1 <-gam(Outstate~Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+Room.Board+Books+ P

gam.m2 <-gam(Outstate~Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+Room.Board+Books+ P

gam.m3 <-gam(Outstate~Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+te(Room.Board, Pers

gam.m3 <-gam(Outstate~Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+te(Room.Board, Pers

gam.m4 <-gam(Outstate~Apps+te(Accept, Enroll)+Top10perc+Top25perc+F.Undergrad+P.Undergrad+te(Room.Board
```

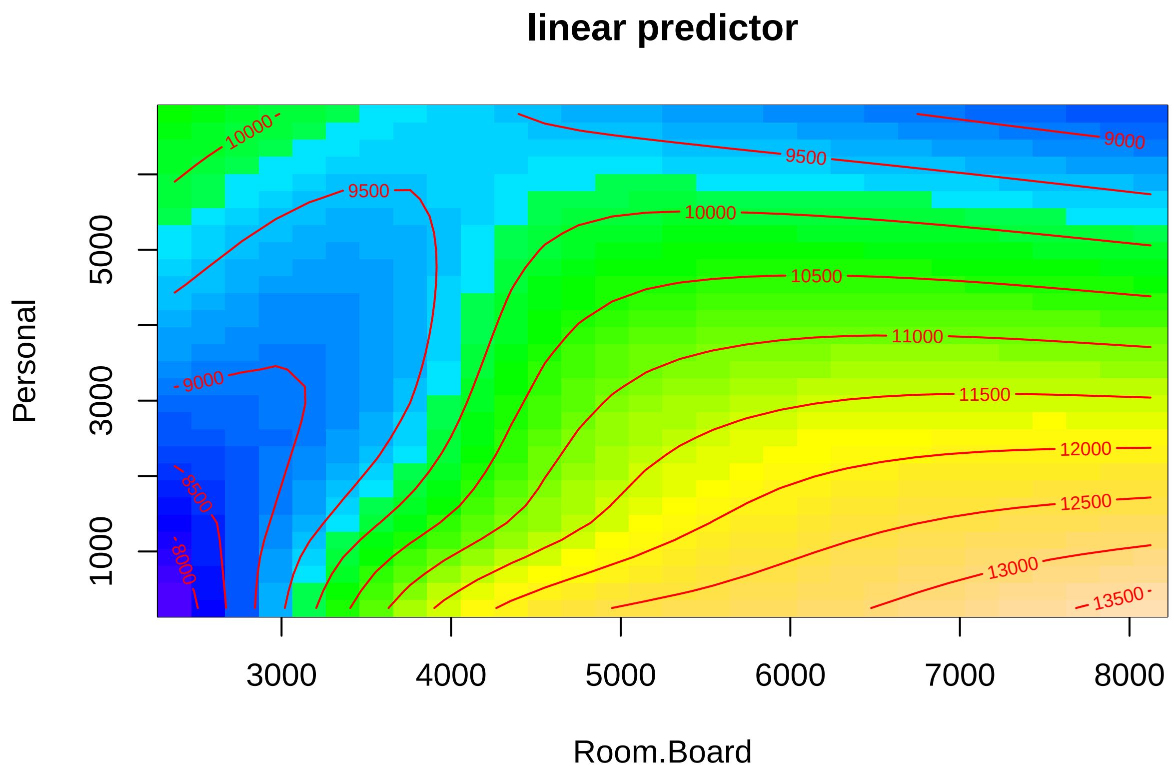
```r
gam.m5 <-gam(Outstate~Apps+te(Accept, Enroll)+te(Top10perc, Top25perc)+s(F.Undergrad)+s(P.Undergrad)+te
```

```
## Warning in term[i] <- attr(terms(reformulate(term[i])), "term.labels"):
## number of items to replace is not a multiple of replacement length
```

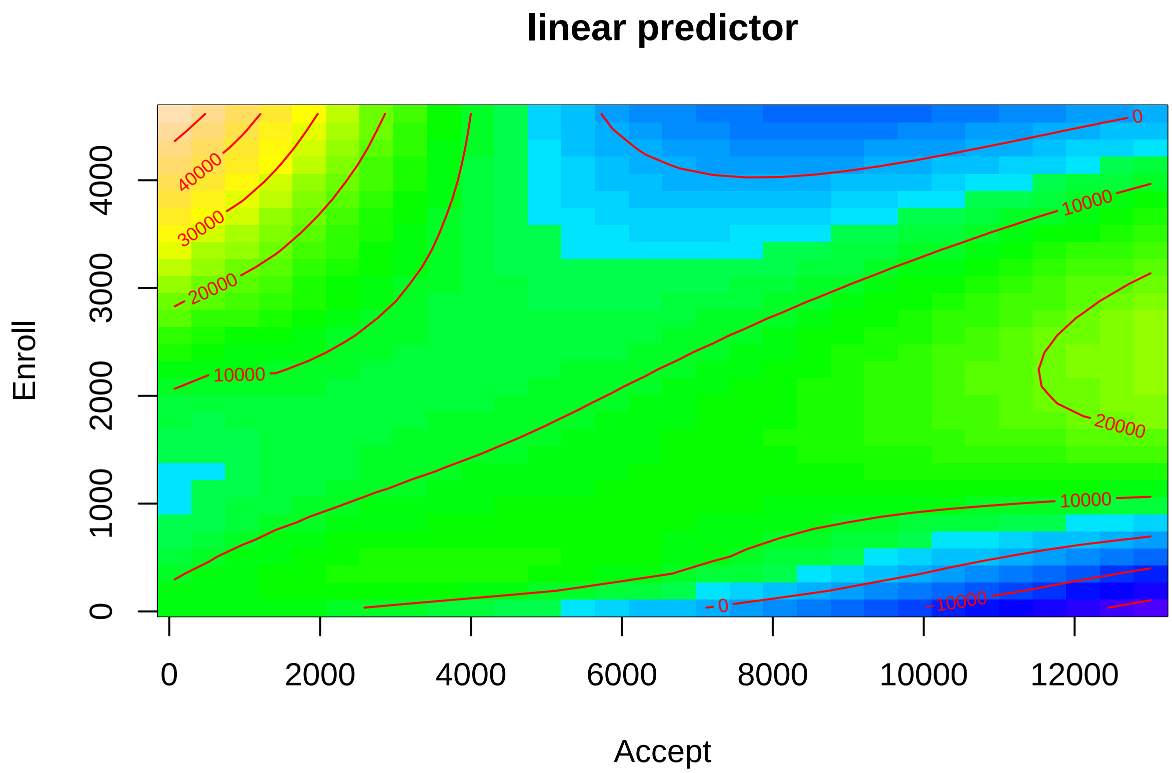
```r
anova(gam.m1, gam.m2, gam.m3, gam.m4, gam.m5, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: Outstate ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 2: Outstate ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + PhD + s(Terminal) +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 3: Outstate ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
##     P.Undergrad + te(Room.Board, Personal) + Books + PhD + s(Terminal) +
##     S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 4: Outstate ~ Apps + te(Accept, Enroll) + Top10perc + Top25perc +
##     F.Undergrad + P.Undergrad + te(Room.Board, Personal) + Books +
##     PhD + s(Terminal) + S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 5: Outstate ~ Apps + te(Accept, Enroll) + te(Top10perc, Top25perc) +
##     s(F.Undergrad) + s(P.Undergrad) + te(Room.Board, Personal) +
##     s(Books) + s(PhD) + s(Terminal) + s(S.F.Ratio) + s(perc.alumni) +
##     s(Expend + Grad.Rate)
##   Resid. Df Resid. Dev      Df  Deviance       F    Pr(>F)
## 1    547.00 2092185295
## 2    542.37 2026858216  4.6295  65327078  5.4324 0.0001158 ***
## 3    534.94 1940535407  7.4277  86322809  4.4741 5.026e-05 ***
## 4    524.29 1810986378 10.6562 129549029  4.6803 1.092e-06 ***
## 5    506.44 1338019137 17.8466 472967241 10.2027 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

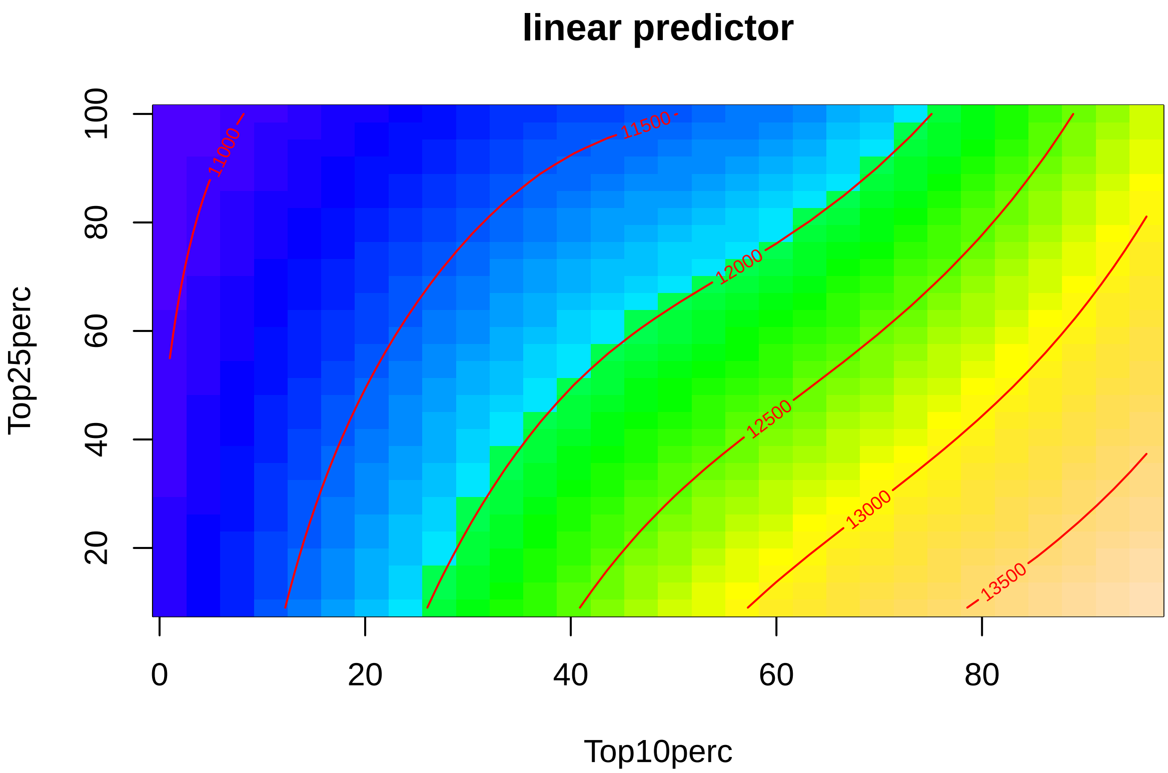
```r
vis.gam(gam.m3, view =c("Room.Board","Personal"),plot.type = "contour", color = "topo")
```



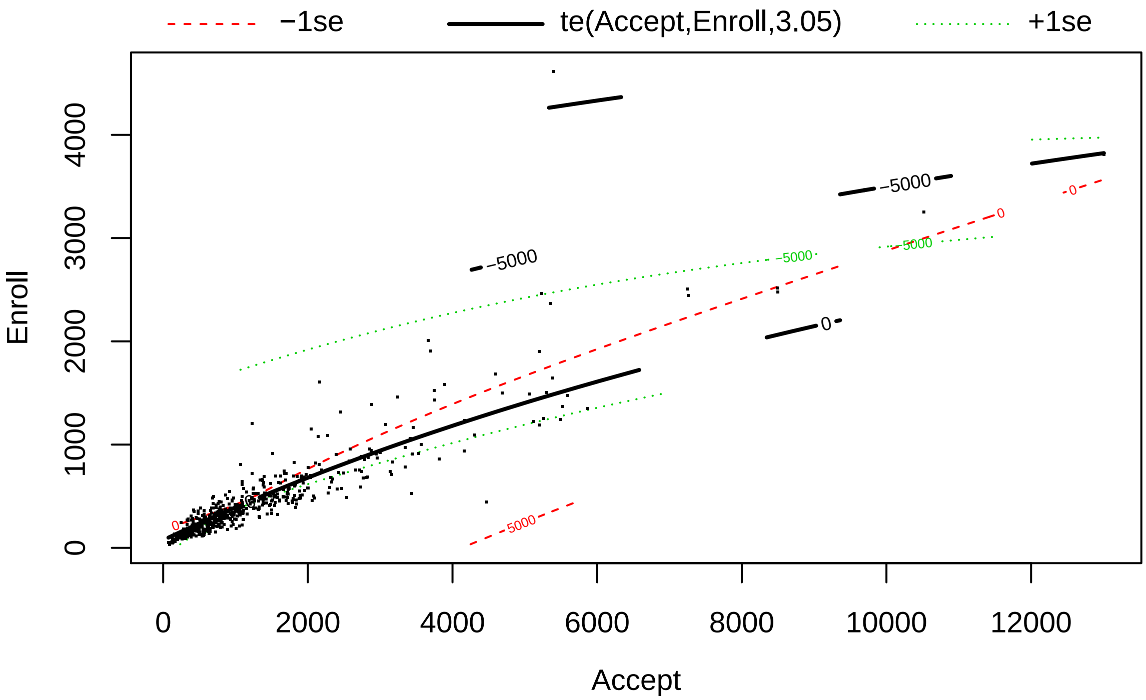

```
vis.gam(gam.m4, view =c("Accept","Enroll"),plot.type = "contour", color = "topo")
```



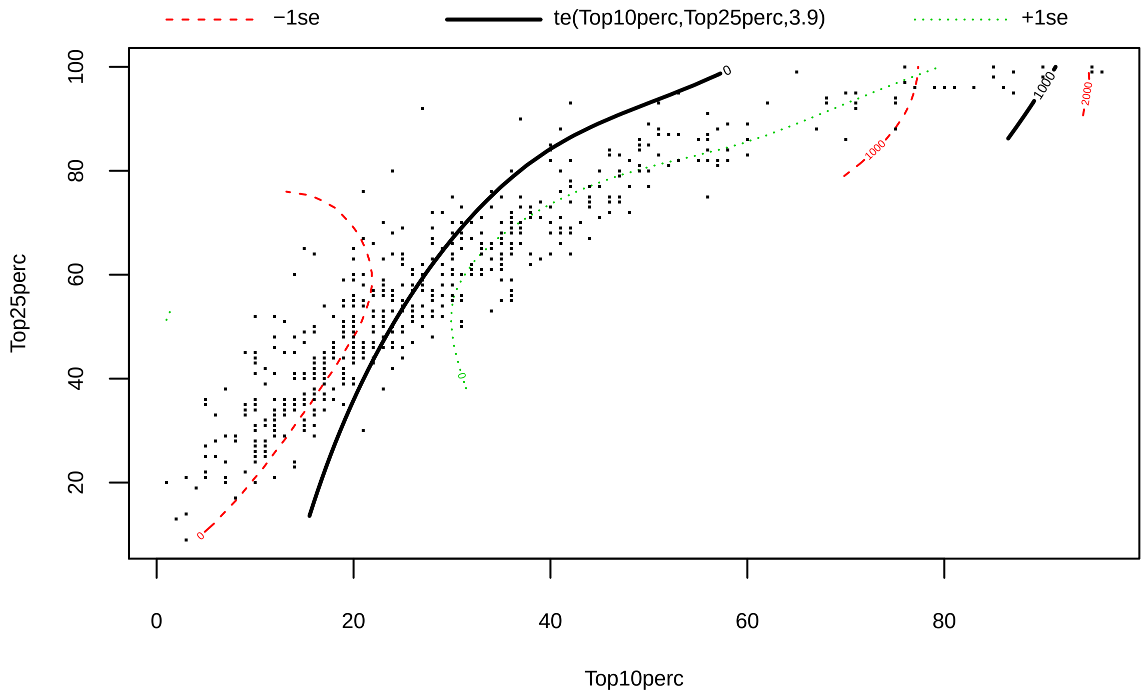

```r
vis.gam(gam.m5, view =c("Top10perc","Top25perc"),plot.type = "contour", color = "topo")
```

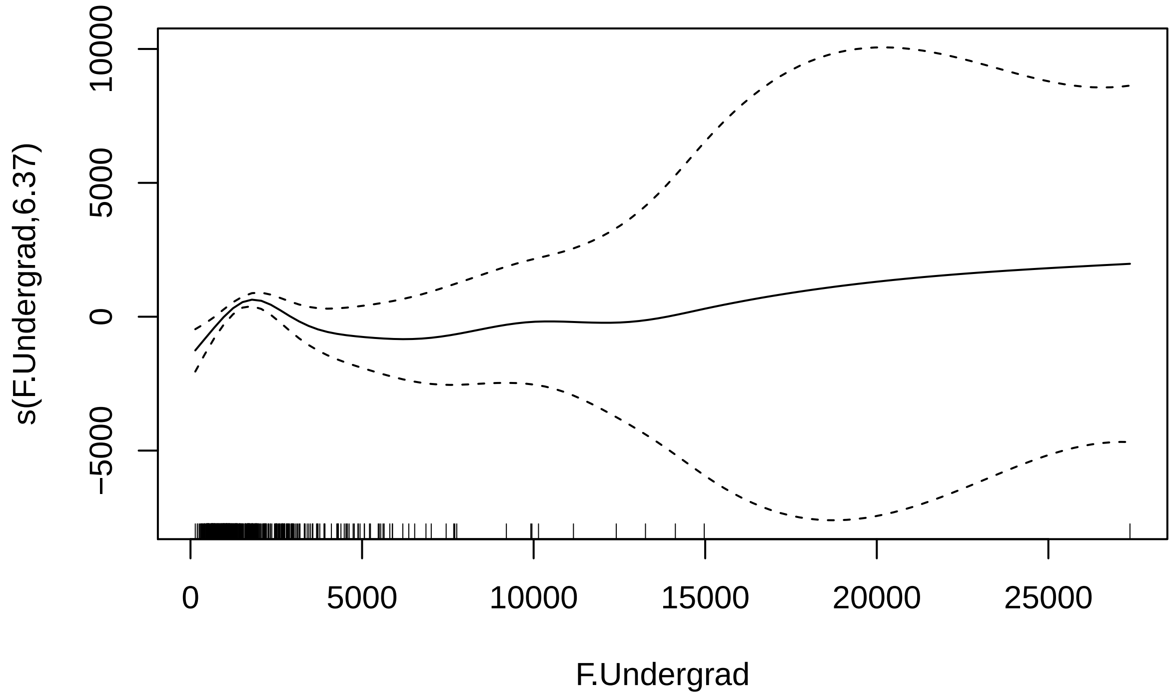

linear predictor
Top25perc
Top10perc
11000
11500
12000
12500
13000
13500
0
20
40
60
80
100

```r
plot(gam.m5)
```

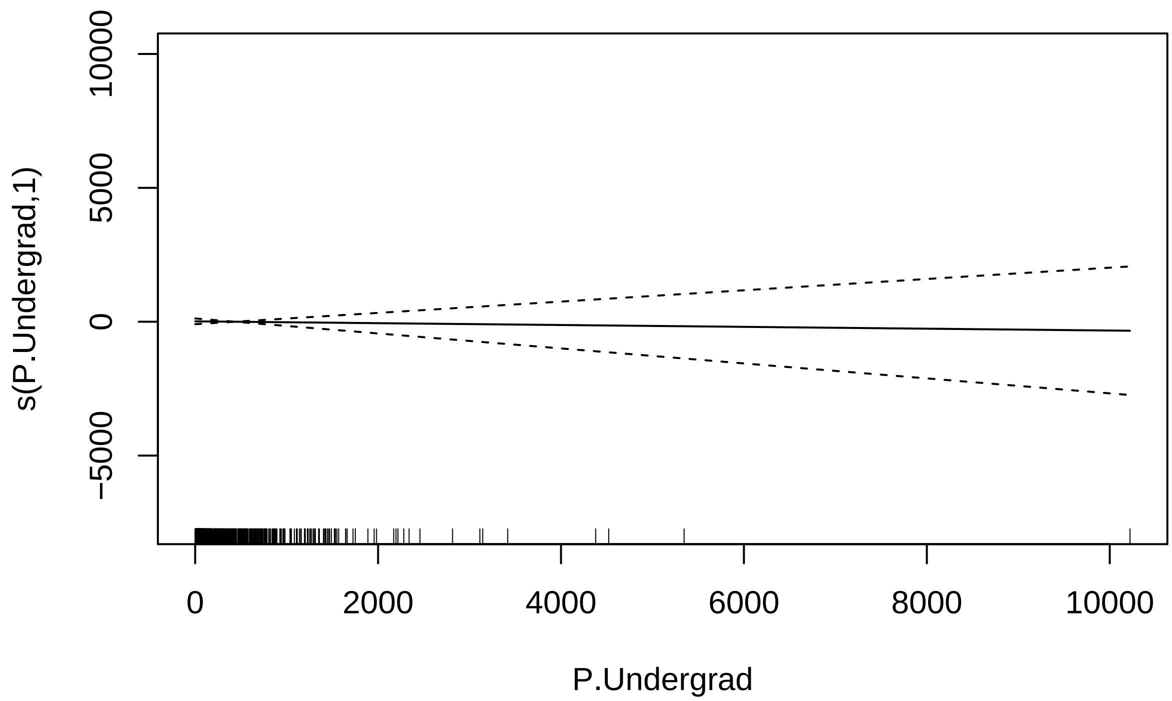


−1se
te(Accept,Enroll,3.05)
+1se
Enroll
Accept
0
1000
2000
3000
4000
0
2000
4000
6000
8000
10000
12000
−5000
−5000
0

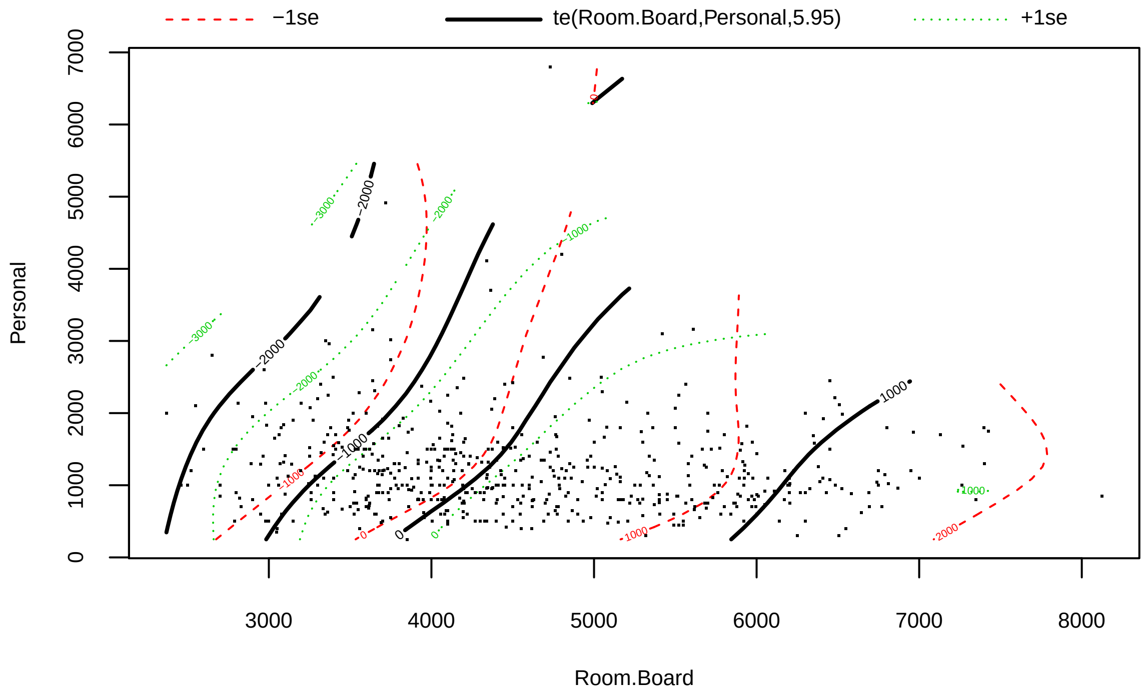

−1se
te(Top10perc,Top25perc,3.9)
+1se
Top25perc
Top10perc
0
20
40
60
80
100
1000
2000

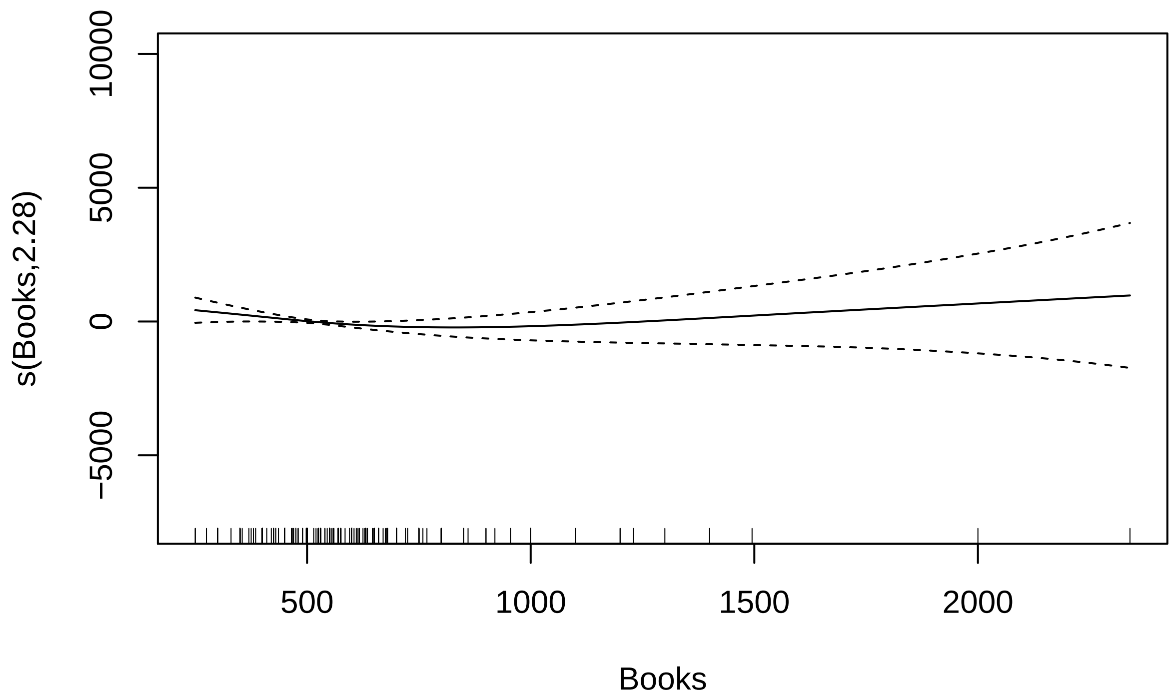


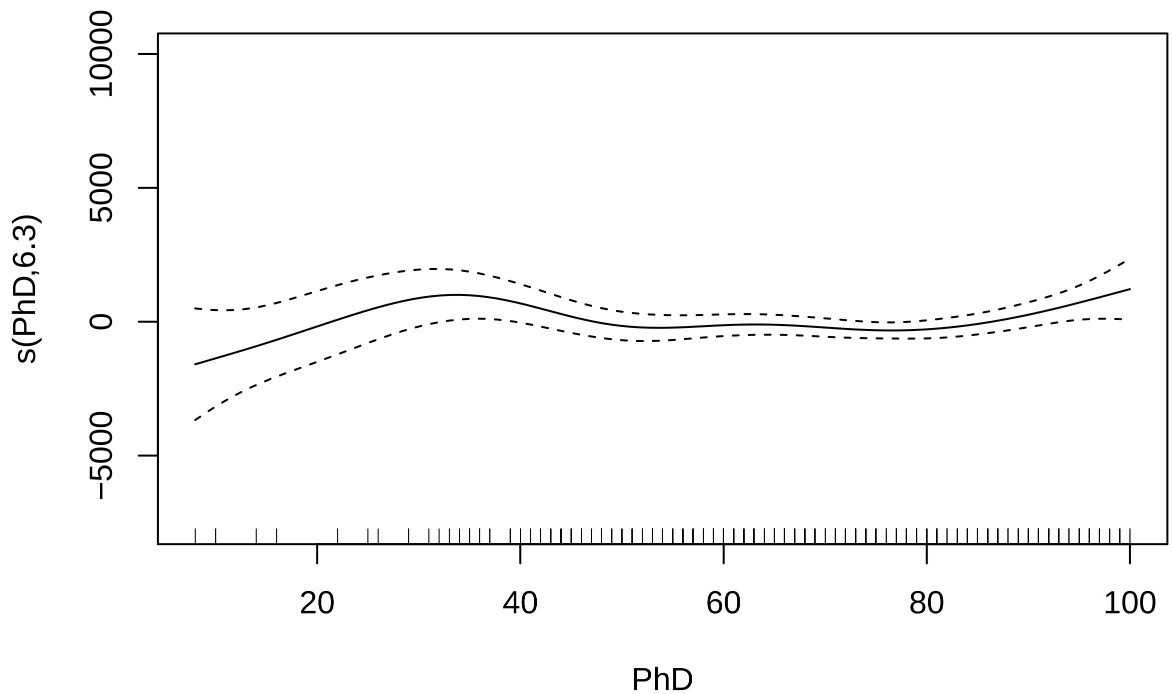
10000
5000
0
−5000
s(F.Undergrad,6.37)
0
5000
10000
15000
20000
25000
F.Undergrad



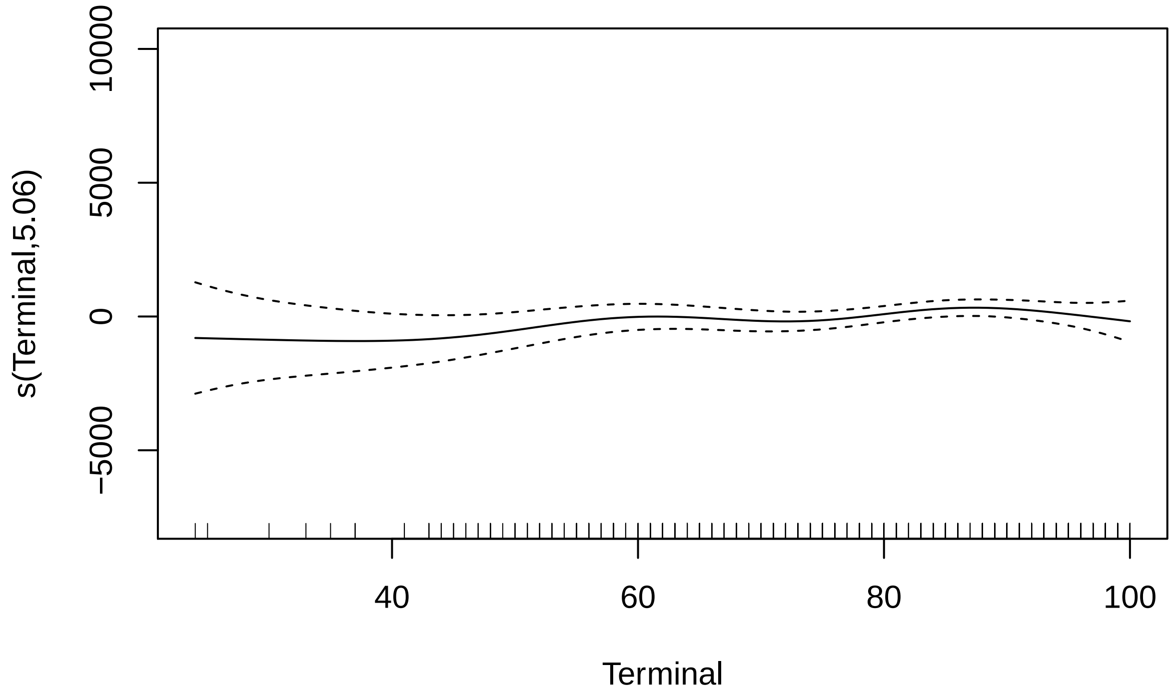
10000
5000
0
−5000
s(P.Undergrad,1)
0
2000
4000
6000
8000
10000
P.Undergrad


−1se
te(Room.Board,Personal,5.95)
+1se
Personal
Room.Board
0
1000
2000
3000
4000
5000
6000
7000
3000
4000
5000
6000
7000
8000

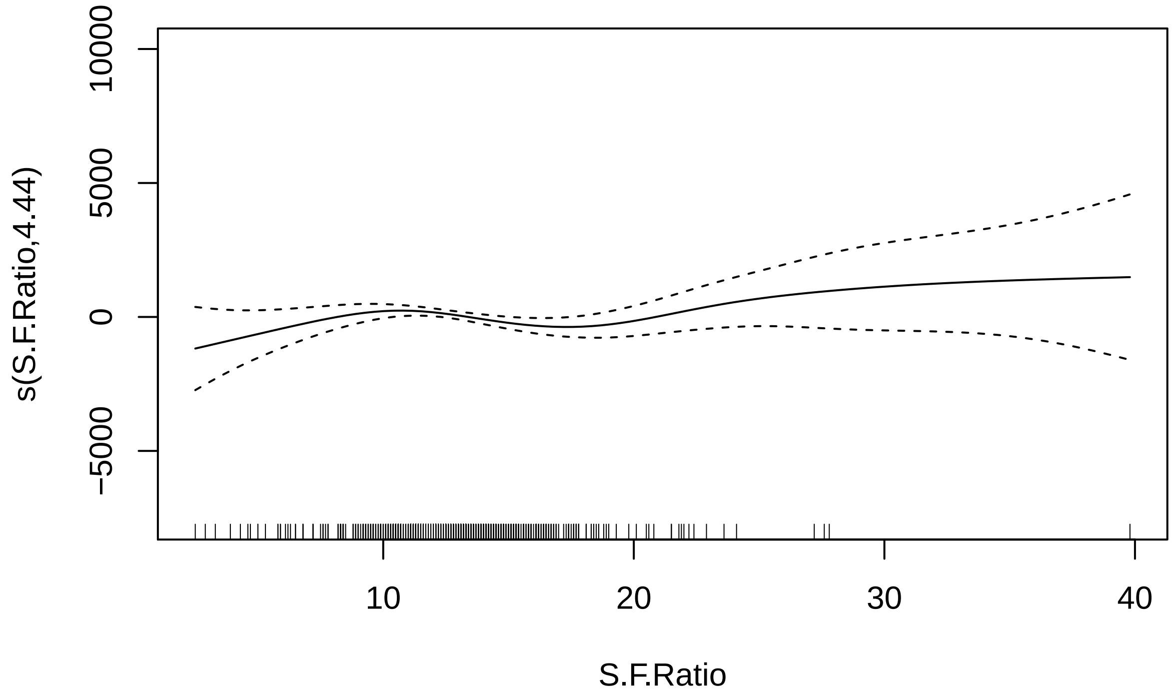

10000
5000
0
−5000
s(Books,2.28)
500
1000
1500
2000
Books

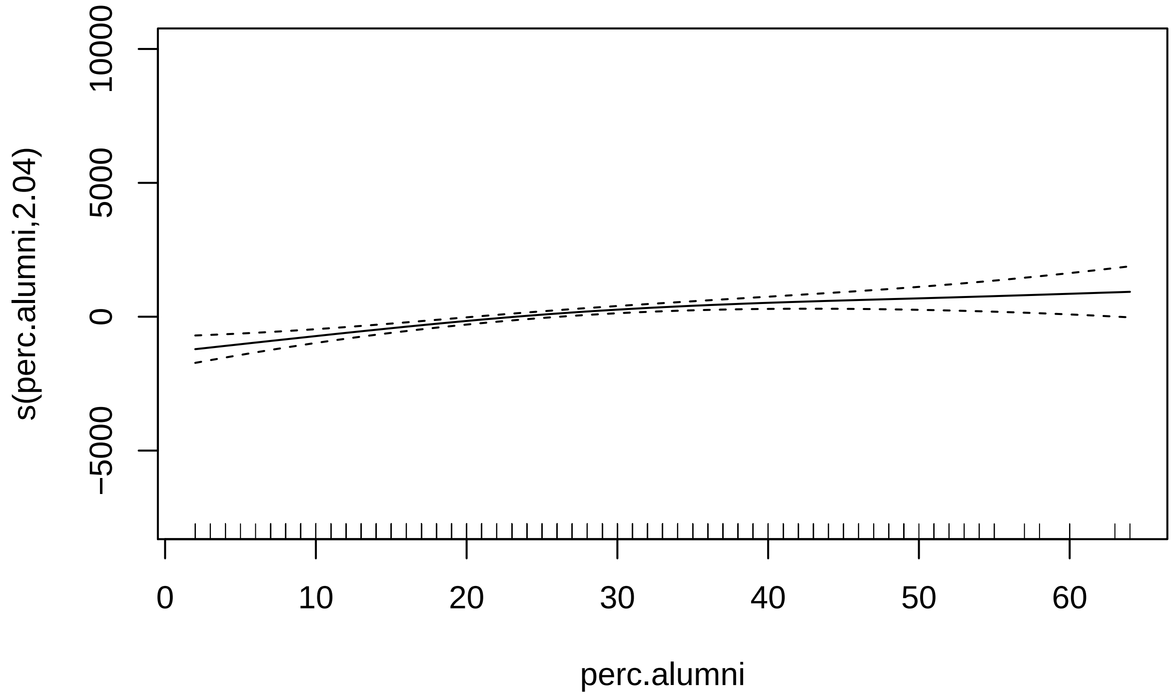


10000
5000
0
−5000
s(PhD,6.3)
20
40
60
80
100
PhD

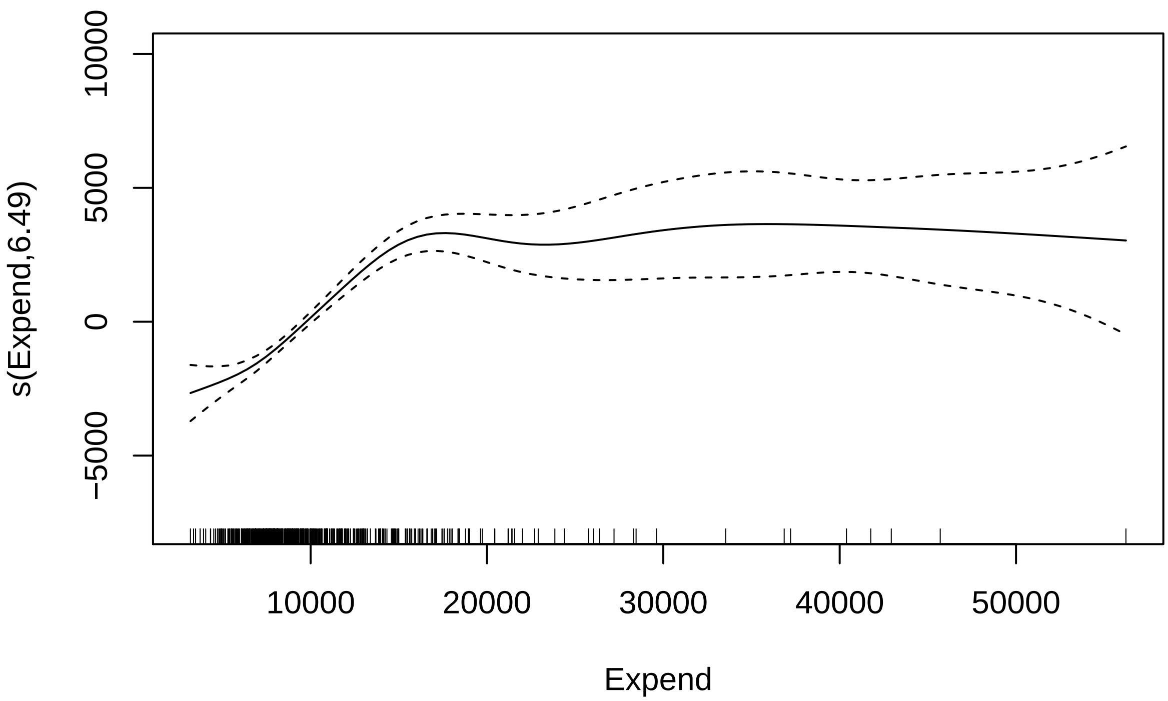


s(Terminal,5.06)
10000
5000
0
−5000
40
60
80
100
Terminal


s(S.F.Ratio,4.44)
10000
5000
0
−5000
10
20
30
40
S.F.Ratio



s(perc.alumni,2.04)
10000
5000
0
−5000
0
10
20
30
40
50
60
perc.alumni



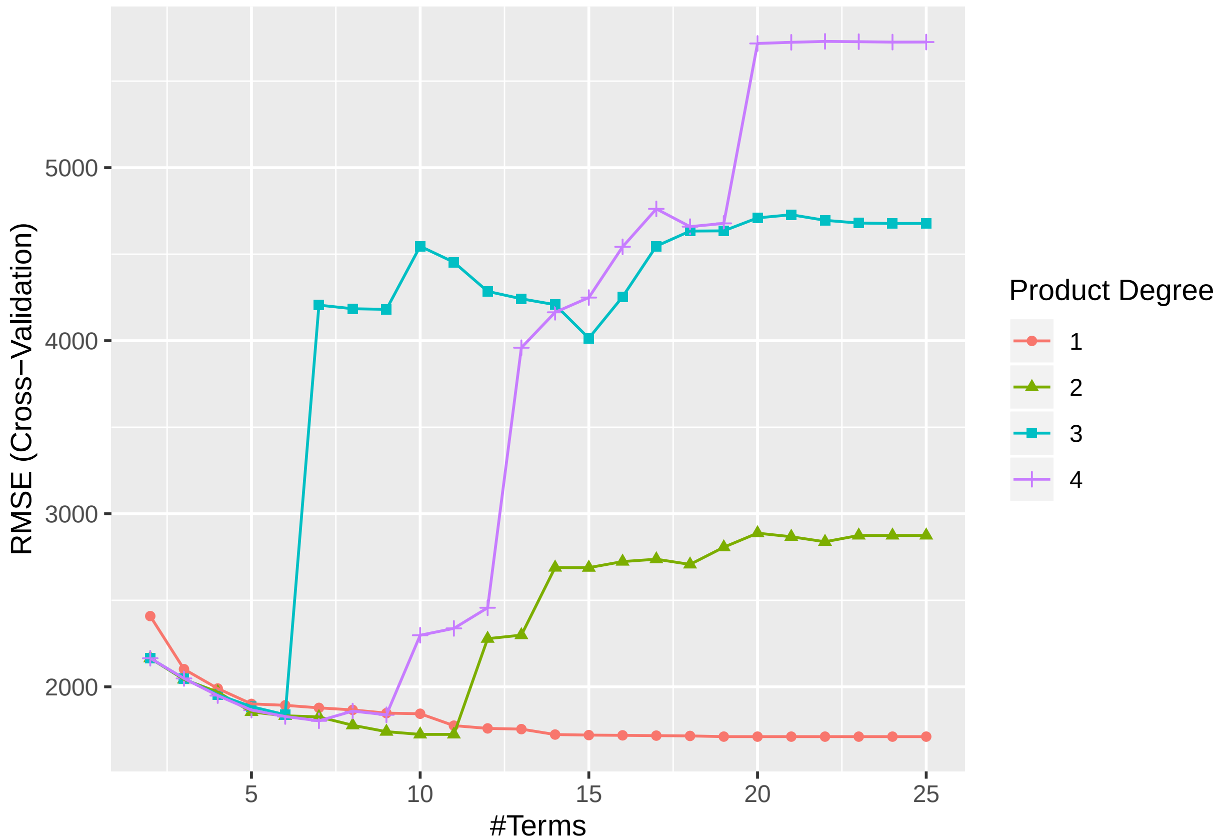
According to the anova procedure, model 5 is the best fitting model. In this model, terminal is still modeled as a nonlinear term, while accept and enroll, as well as room_board/personal and top10/top25 are modeled as tensors. Also, all other terms were modeled nonlinearly with the s() function. These terms, are shown to be the best fitting model compared to model 1, 2, 3, and 4. As I stated in the previous question, with the number of staff with a terminal degree, there appears to be no real affect until about the 70th percentile. After this point tuition begins to rise rapidly. I am not sure how to interpret the tensors. Increased room and board appears to affect out of state tuition more noticably when estimated personal spending is low. PHD's also appears to be nonlinear, tuition rises quickly at first as PHD proportion rises, stays flat, and then rises again once >90% of faculty have PHD's

There are possibly better models still. It would be nice to select variables that should be modeled as nonlinear terms or tensors automatically using the MGCV package.

(d) Fit a multivariate adaptive regression spline (MARS) model using all the predictors. Report the final model. Present the partial dependence plot of an arbitrary predictor in your final model.

```
ctrl1 <-trainControl(method = "cv", number = 10)
mars_grid <-expand.grid(degree = 1:4,
                        nprune =2:25)

mars.fit <-train(x, y,
                 method = "earth",
```

```r
                  tuneGrid = mars_grid,
                  trControl = ctrl1)

ggplot(mars.fit)
```




```r
mars.fit$bestTune
```

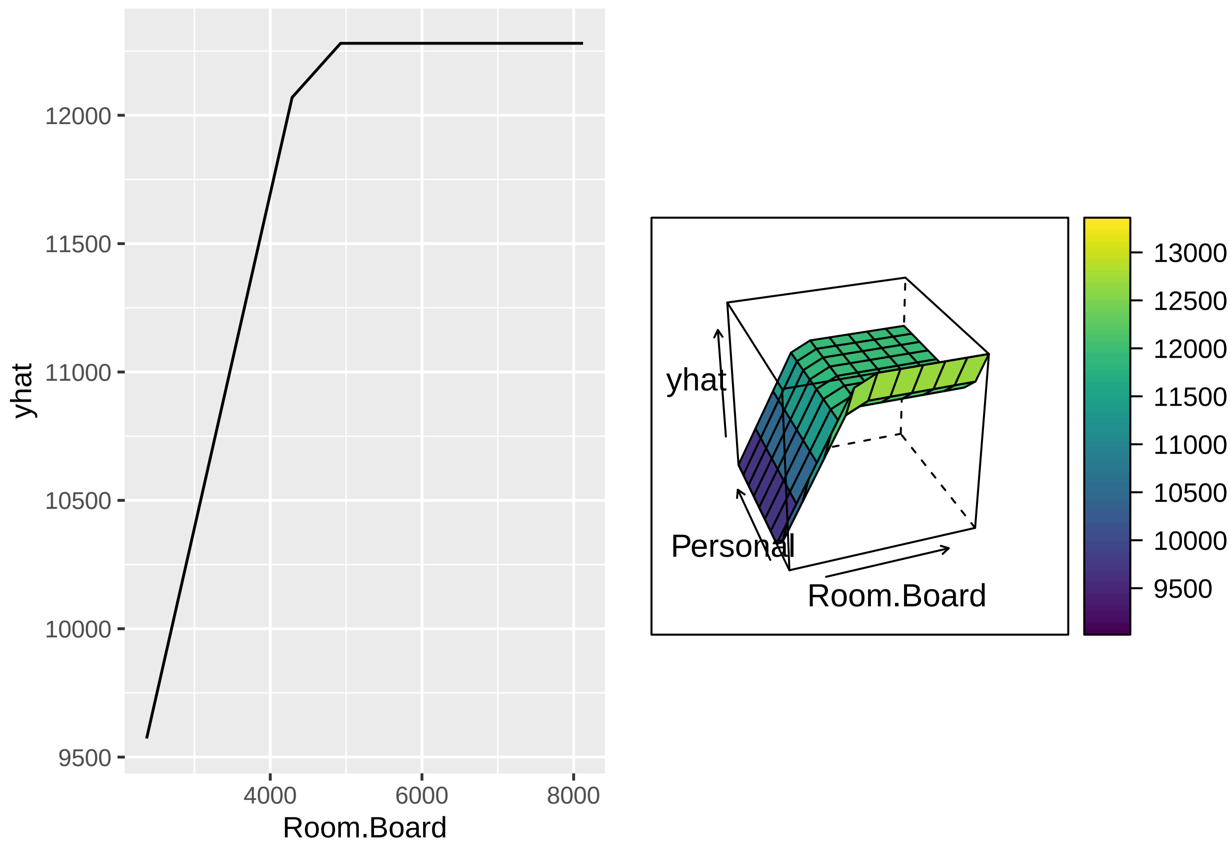
```
##    nprune degree
## 19     20      1
```

```r
coef(mars.fit$finalModel)
```

```
##          (Intercept)     h(Expend-15365)  h(4450-Room.Board)
##        11263.1644592          -0.6232067          -1.3019686
##      h(Grad.Rate-97)     h(97-Grad.Rate) h(F.Undergrad-1355)
##         -240.4181519         -22.0705355          -0.3455994
## h(1355-F.Undergrad)   h(22-perc.alumni)        h(Apps-3712)
##           -1.6975458         -74.5653989           7.0546188
##     h(1300-Personal)        h(913-Enroll)      h(2193-Accept)
##            1.0518029           4.7810367          -1.8020047
##       h(Expend-6881)        h(Apps-3877)    h(S.F.Ratio-10.1)
##            0.6430814          -6.7165025        -131.6442126
##    h(10.1-S.F.Ratio)   h(S.F.Ratio-17.8)          h(PhD-81)
##         -276.0954428         264.3799299          58.2239473
```

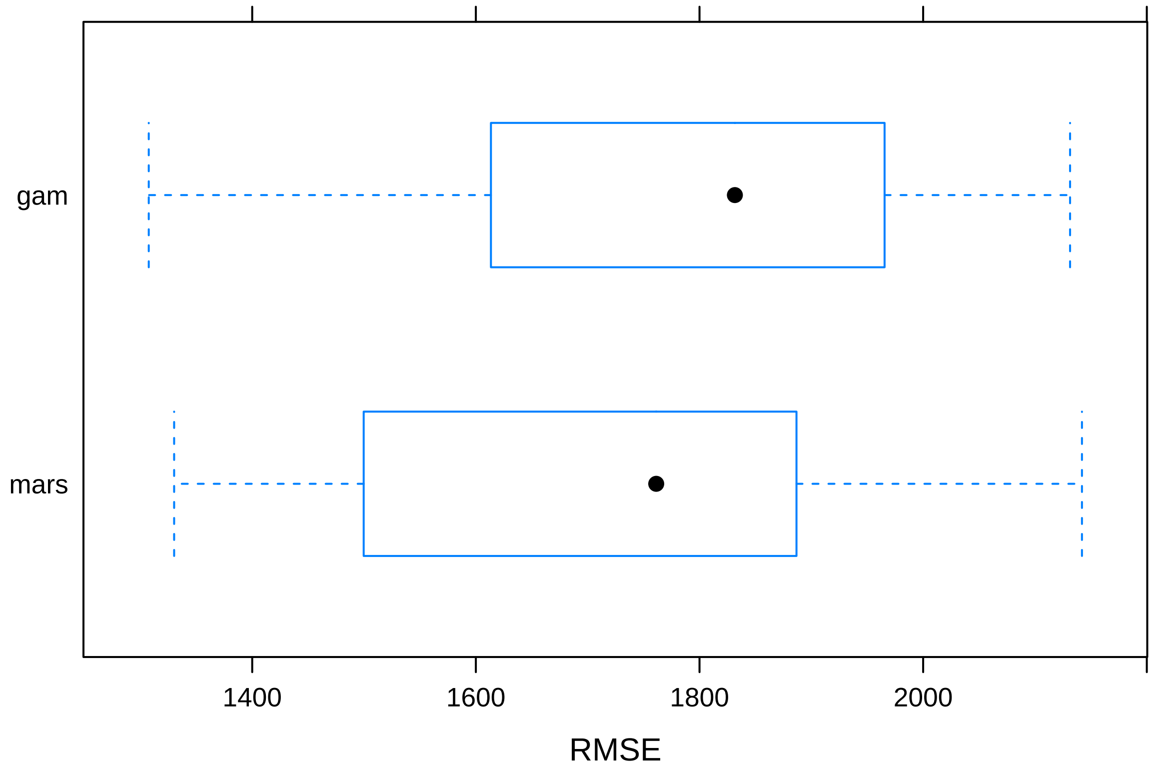
Partial Dependence plot

```
p1 <-partial(mars.fit, pred.var =c("Room.Board"), grid.resolution = 10)%>% autoplot()

p2 <-partial(mars.fit, pred.var =c("Room.Board","Personal"), grid.resolution = 10)%>%plotPartial(levelpl

grid.arrange(p1, p2, ncol = 2)
```



Compare to caret GAM

```
gam.fit <-train(x, y,
                method = "gam",
                tuneGrid =data.frame(method = "GCV.Cp", select =c(TRUE,FALSE)),
                trControl = ctrl1)


bwplot(resamples(list(mars = mars.fit,gam = gam.fit)), metric = "RMSE")
```




```r
resamp <- resamples(list(gam=gam.fit,
                         mars=mars.fit))

summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: gam, mars
## Number of resamples: 10
##
## MAE
##          Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## gam  1012.660 1251.970 1440.060 1384.288 1516.381 1627.591    0
## mars 1062.586 1159.266 1385.627 1334.407 1438.803 1637.302    0
##
## RMSE
##          Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## gam  1307.512 1652.394 1831.630 1788.109 1951.456 2131.387    0
## mars 1330.183 1517.869 1761.304 1712.300 1877.553 2142.005    0
##
## Rsquared
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## gam  0.6757425 0.7425273 0.7588392 0.7715996 0.8122629 0.8862215    0
## mars 0.6889092 0.7297082 0.8039105 0.7869848 0.8272647 0.8875854    0
```

This step was not asked for, but the mars fit appears to have a slightly lower RMSE, but this is comparing the Caret gam and mars models, not my final model using the GAM package.

(e) Based on the above GAM and MARS models, predict the out-of-state tuition of Columbia University.

```
columbia<-college %>% filter(College=="Columbia University")
pred.gam <-predict(gam.fit, newdata =columbia)
pred.mars <-predict(mars.fit, newdata =columbia)

#predicted using the GAM package instead of Caret
pred.gam.m5=predict(gam.m5,newdata=columbia,se = TRUE)


pred.gam
```

```
##        1
## 17728.51
```

```
pred.mars
```

```
##             y
## [1,] 18456.89
```

```
pred.gam.m5
```

```
## $fit
##        1
## 17929.48
##
## $se.fit
##        1
## 798.5085
```

The predicted value for the caret GAM model is 17728.51 dollars, and the predicted value for the MARs model is $18456.89 The predicted value using the fifth GAM package model was 17929.48