homework2

Na Yun Cho

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(splines)
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

```
library(pdp)
library(earth)
```

```
## Loading required package: Formula
```

```
## Loading required package: plotmo
```

```
## Loading required package: plotrix
```

```
## Loading required package: TeachingDemos
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

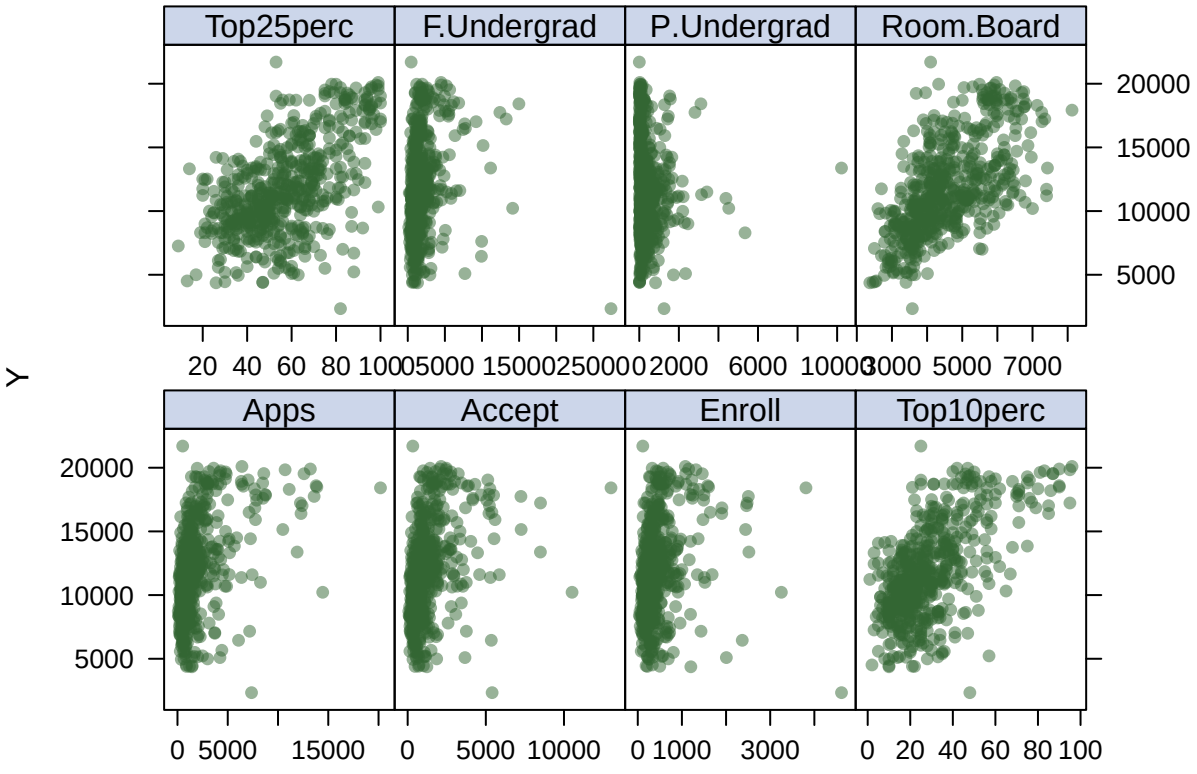
```
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

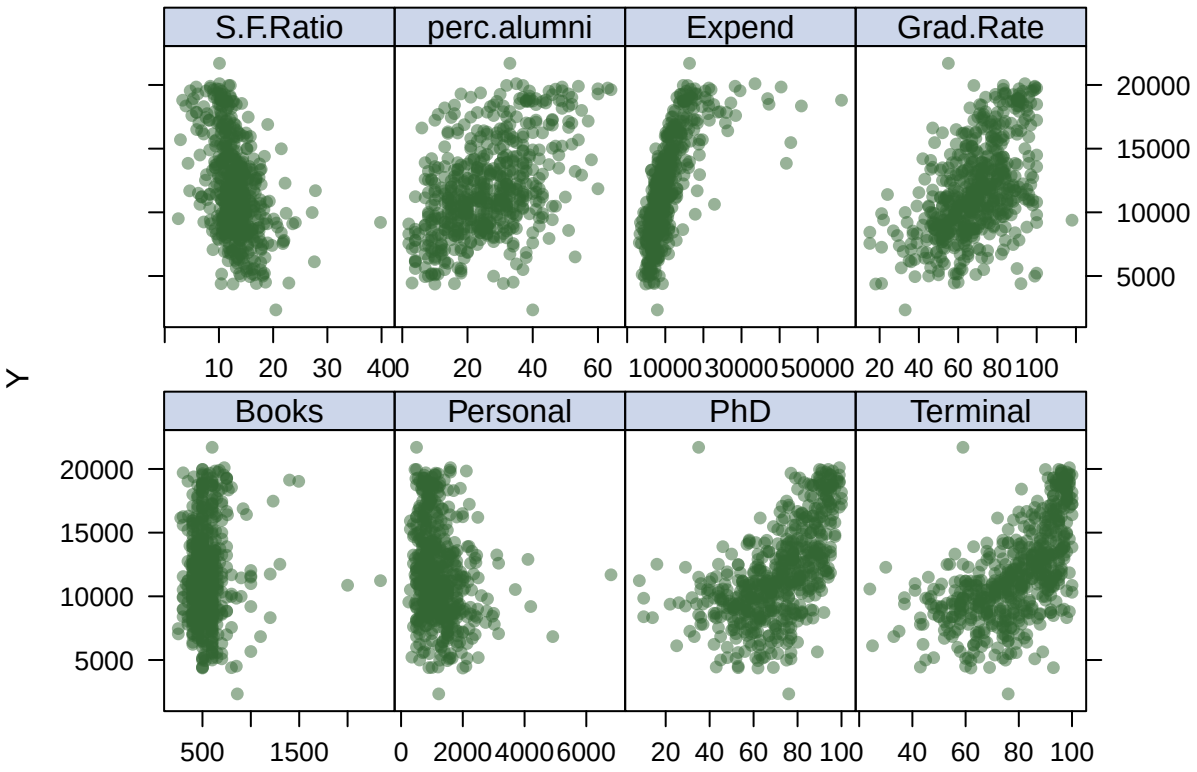
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::collapse() masks nlme::collapse()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x purrr::lift()     masks caret::lift()
## x purrr::partial()  masks pdp::partial()
```

```
library(ggplot2)
```

(a) Exploratory data analysis

```
college = read.csv("./data/data.csv")
college1 <- college[-125,]

college2 = data.matrix(college1, rownames.force = NA)
x <- college2 [ ,-c(1,9)]
y <- college2 [ , 9]


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))
```





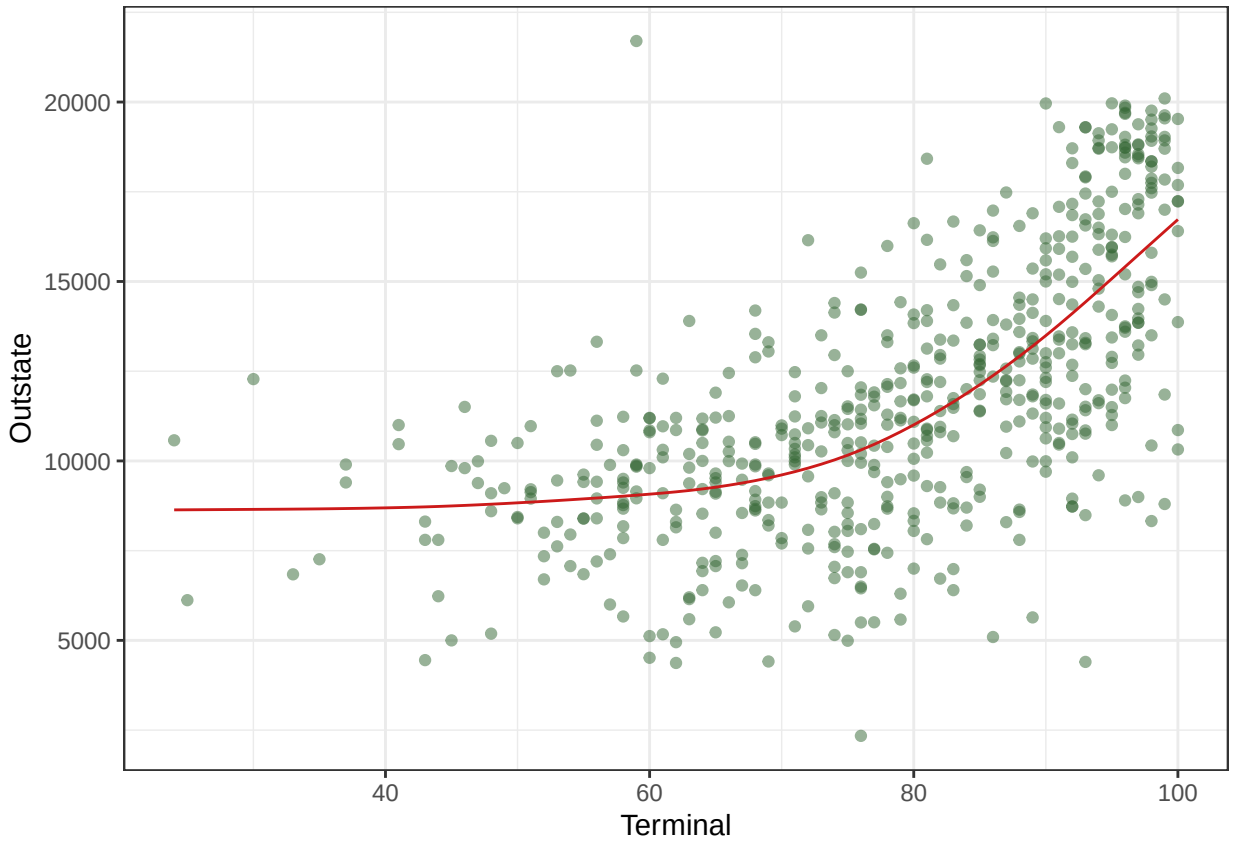
Interpretation: From this exploratory data analysis, I could see that the predictors 'F.Undergrad', 'P.Undergrad', 'Apps', 'Accept', 'Enroll', 'Terminal', and 'Books' show a relatively non-linear trend compared to other predictors. The predictors 'Top25perc', 'Room.Board', 'Top10perc', 'perc.alumni', 'Grad.Rate', 'Expend', and 'PhD' showed a generally increasing trend that looks quite linear. On the other hand, 'S.F.Ratio' and 'Personal' seemed to show a slightly decreasing trend that is quite linear. To check the associations of each predictor with the outcome 'Outstate' in more detail, further analyses would have to be done.

(b) Fit a smoothing spline model using 'Terminal' as the only predictor

```
# using GCV method
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate)
fit.ss$df
```

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

```
Terminallims <- range(college1$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= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



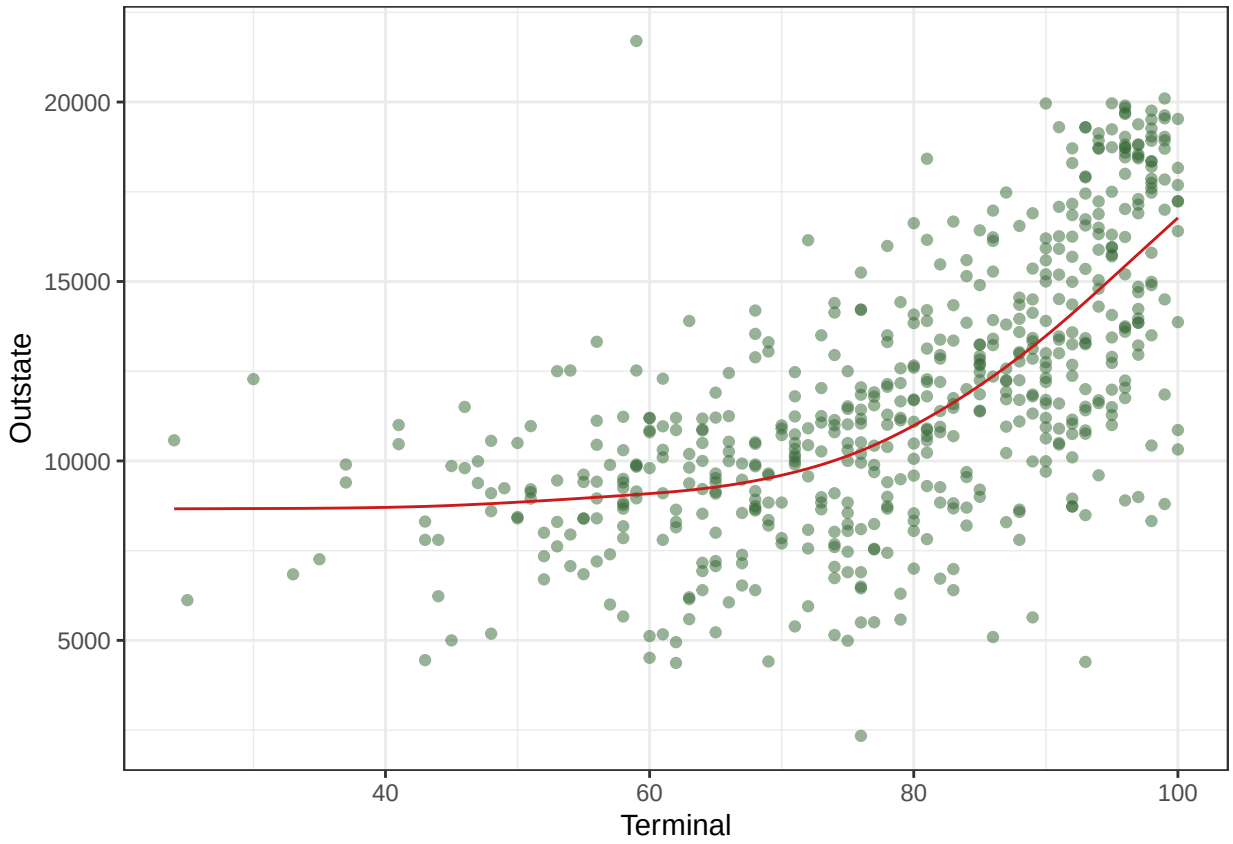

```
#Using LOOCV method
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, cv = TRUE)
```

```
## Warning in smooth.spline(college1$Terminal, college1$Outstate, cv = TRUE):
## cross-validation with non-unique 'x' values seems doubtful
```

```
fit.ss$df
```

```
## [1] 4.686019
```

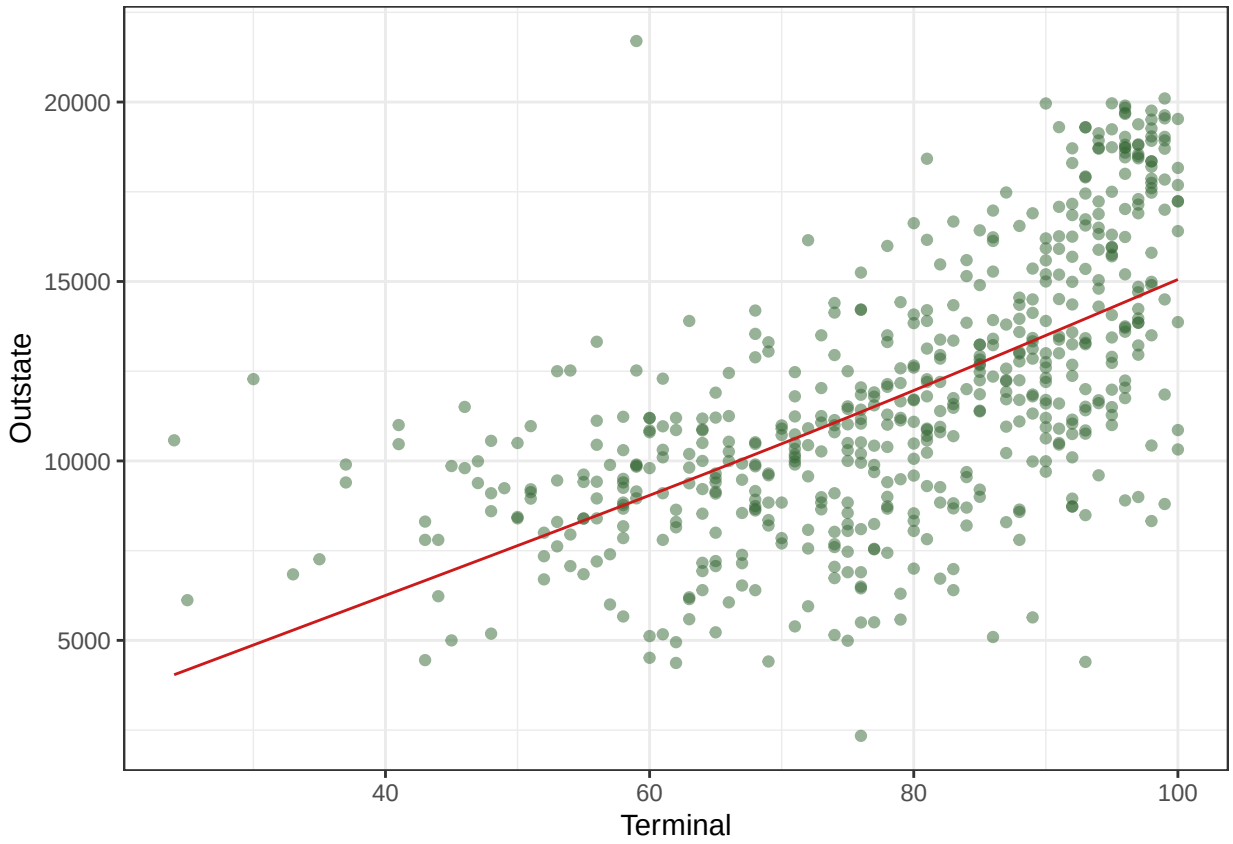
```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

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

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```




```
#Using arbitrary lambda values
#Using lambda = 10
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=10)
fit.ss$df
```

```
## [1] 2.06511
```

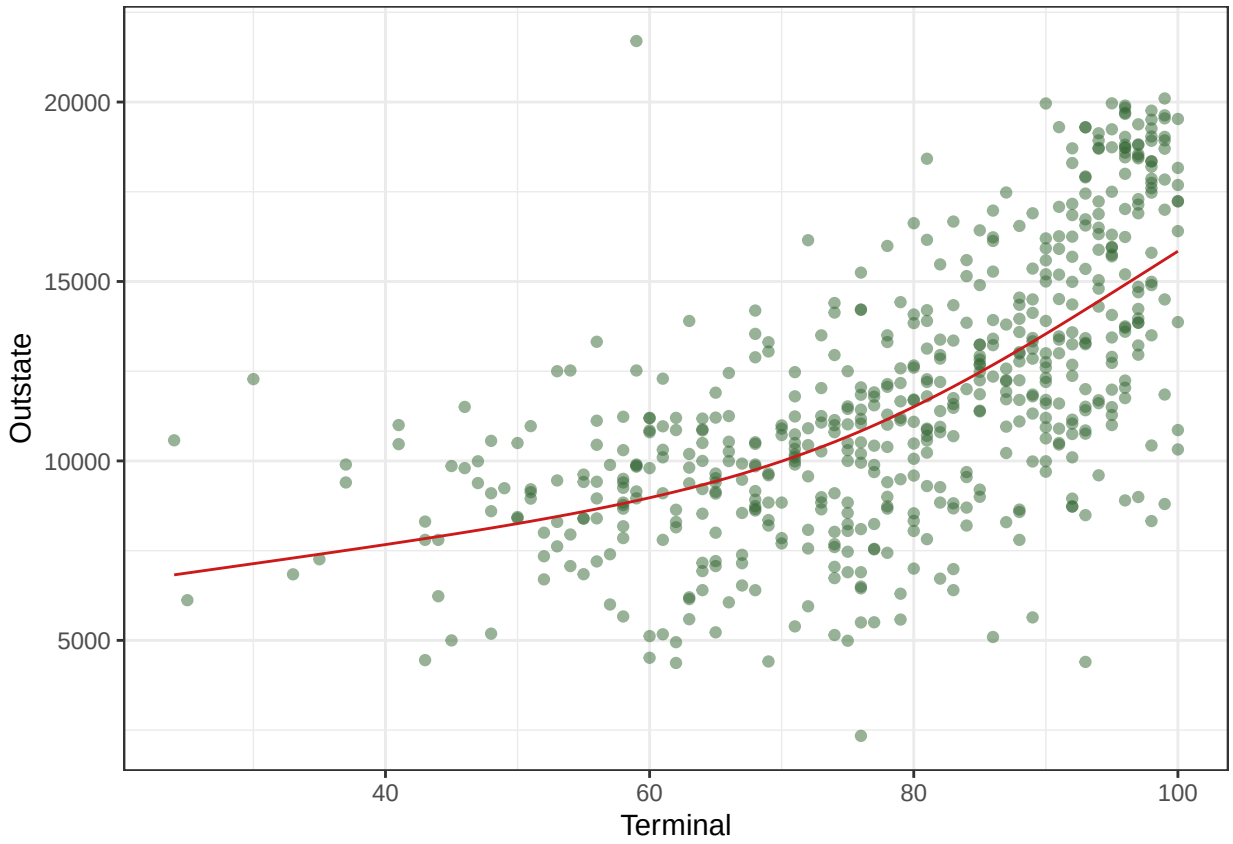
```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

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

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```




```
#Using lambda = 0.5
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=0.5)
fit.ss$df
```

```
## [1] 2.761186
```

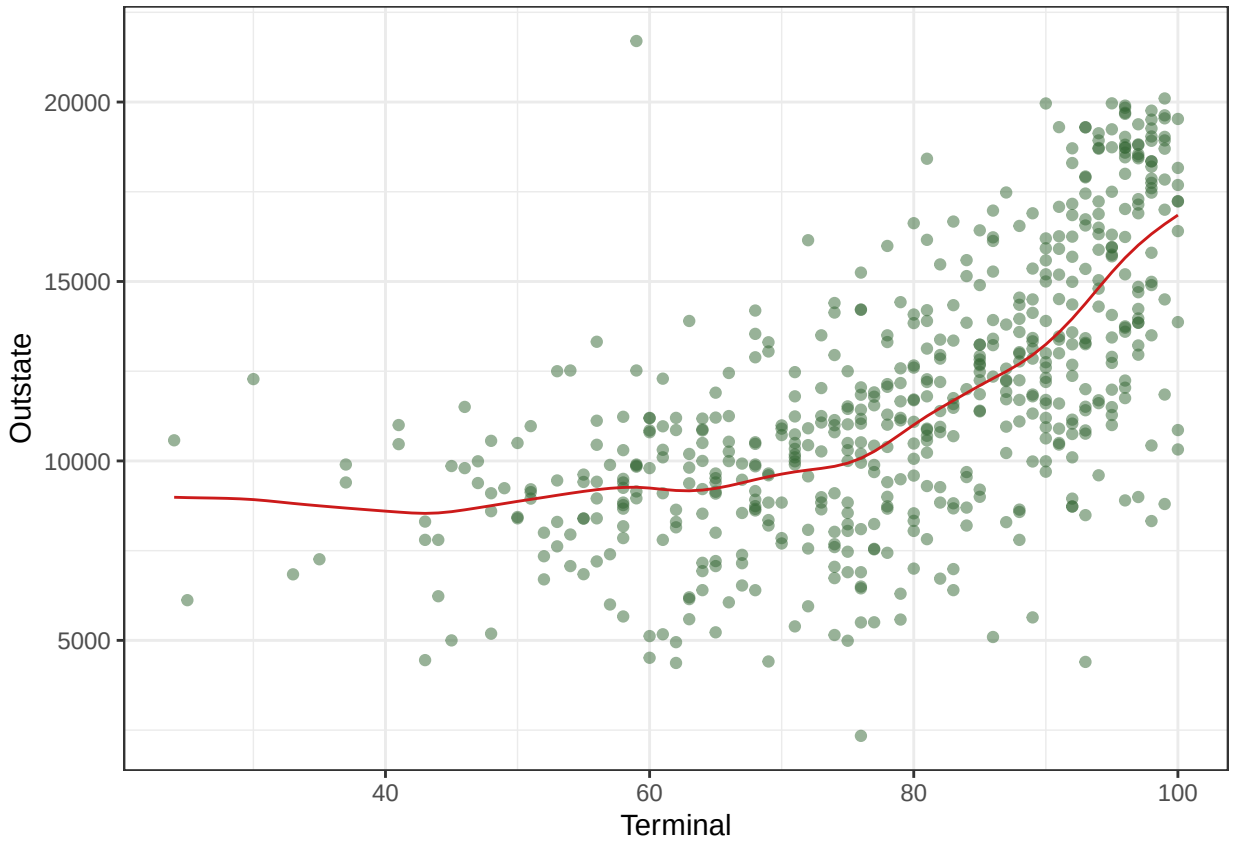
```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

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

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5)
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```




```
#Using lambda = 0.001
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=0.001)
fit.ss$df
```

```
## [1] 9.838879
```

```
Terminallims <- range(college1$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= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5)
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



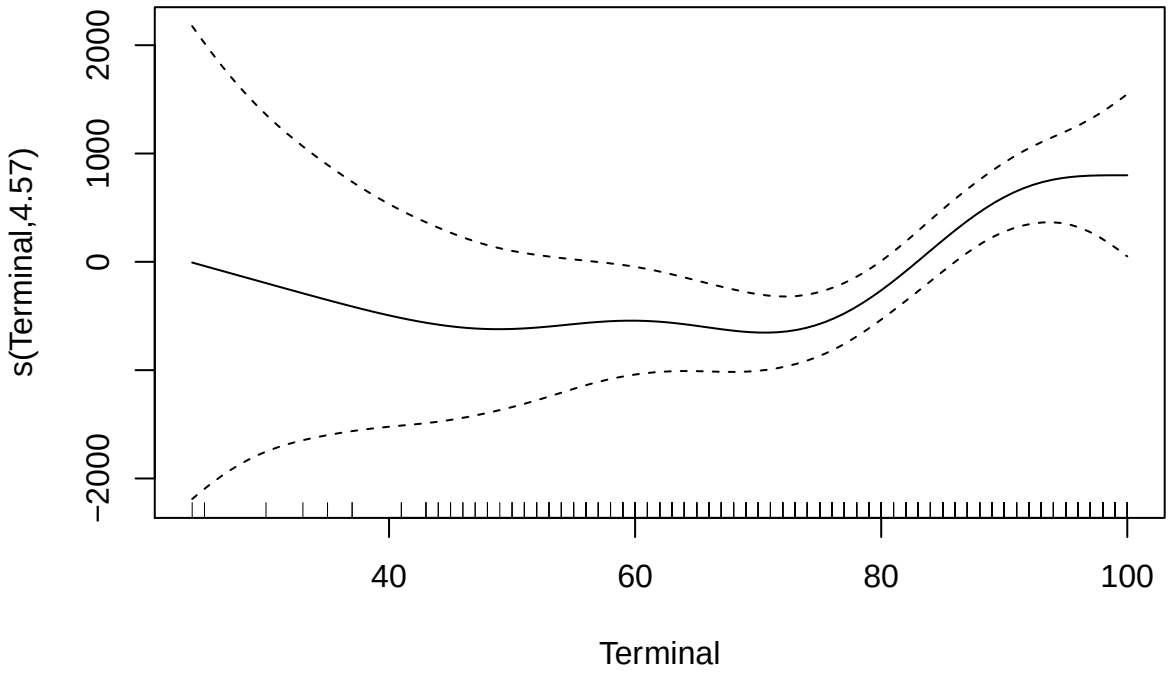
(c) Fit a GAM model using all the predictors

```
gam.m1 <- gam(Outstate~ Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+Room.Board+Books
              +Personal+Terminal+PhD+S.F.Ratio+perc.alumni+Expend+Grad.Rate, data = college1)

gam.m2 <- gam(Outstate~ Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+Room.Board+Books
              +Personal+s(Terminal)+PhD+S.F.Ratio+perc.alumni+Expend+Grad.Rate, data = college1)

gam.m3 <- gam(Outstate~ s(Apps)+s(Accept)+s(Enroll)+Top10perc+Top25perc+s(F.Undergrad)+s(P.Undergrad)
              +Room.Board+s(Books)+Personal+s(Terminal)+PhD+S.F.Ratio+perc.alumni+Expend+Grad.Rate,
              data = college1)

gam.m4 <- gam(Outstate~ s(Apps)+s(Accept)+ s(Enroll)+Top10perc+Top25perc+te(F.Undergrad,P.Undergrad)
              +Room.Board+s(Books)+Personal+s(Terminal)+PhD+S.F.Ratio+perc.alumni+Expend+Grad.Rate,
              data = college1)

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

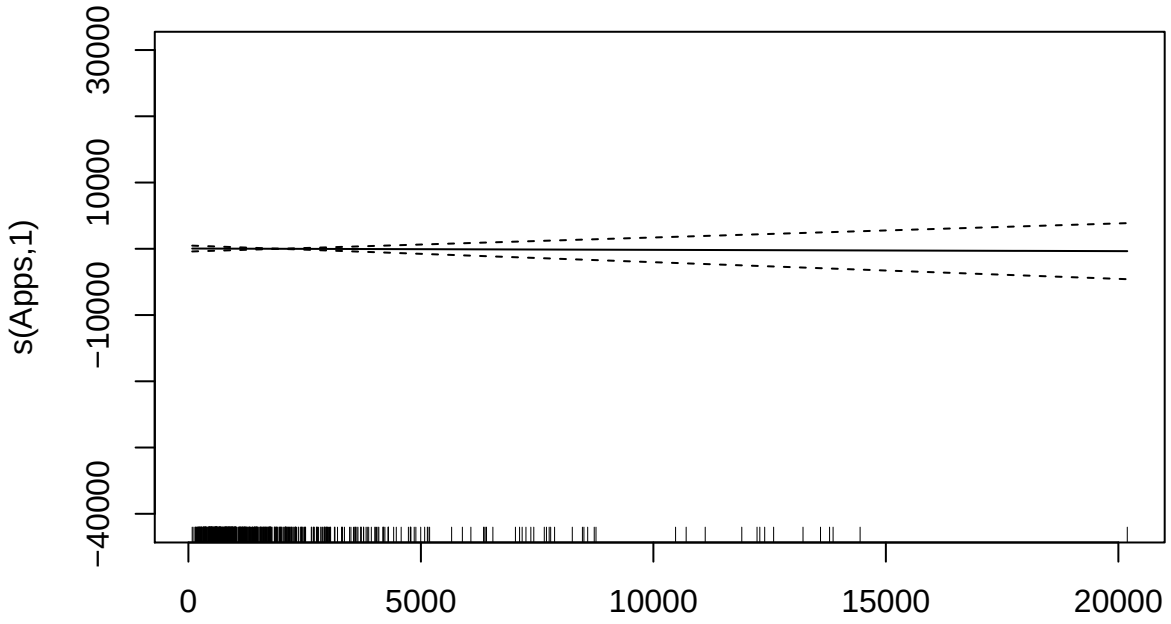
```
## Analysis of Deviance Table
##
## Model 1: Outstate ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
##     P.Undergrad + Room.Board + Books + Personal + Terminal +
##     PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 2: Outstate ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad +
```

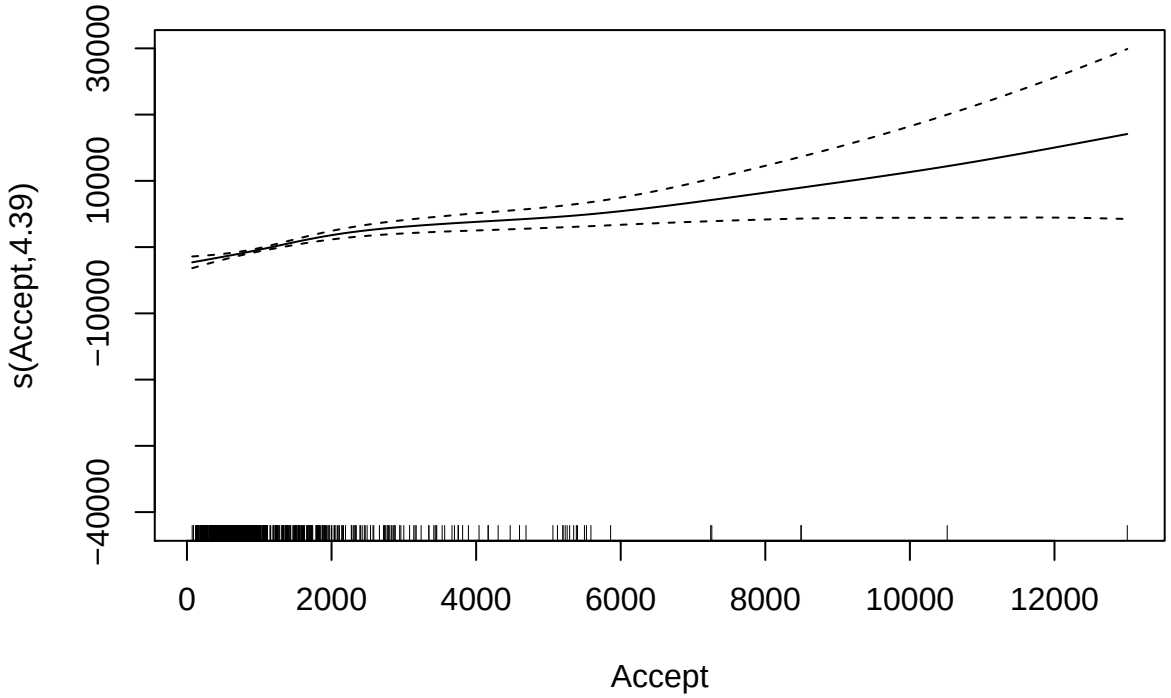
```
##     P.Undergrad + Room.Board + Books + Personal + s(Terminal) +
##     PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate
## Model 3: Outstate ~ s(Apps) + s(Accept) + s(Enroll) + Top10perc + Top25perc +
##     s(F.Undergrad) + s(P.Undergrad) + Room.Board + s(Books) +
##     Personal + s(Terminal) + PhD + S.F.Ratio + perc.alumni +
##     Expend + Grad.Rate
## Model 4: Outstate ~ s(Apps) + s(Accept) + s(Enroll) + Top10perc + Top25perc +
##     te(F.Undergrad, P.Undergrad) + Room.Board + s(Books) + Personal +
##     s(Terminal) + PhD + S.F.Ratio + perc.alumni + Expend + Grad.Rate
##   Resid. Df Resid. Dev      Df  Deviance      F    Pr(>F)
## 1    547.00 2092185295
## 2    542.37 2026858216  4.6295  65327078 4.1413  0.001481 **
## 3    527.63 1829988934 14.7408 196869282 3.9195 1.244e-06 ***
## 4    521.24 1793985069  6.3874  36003865 1.6543  0.125309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

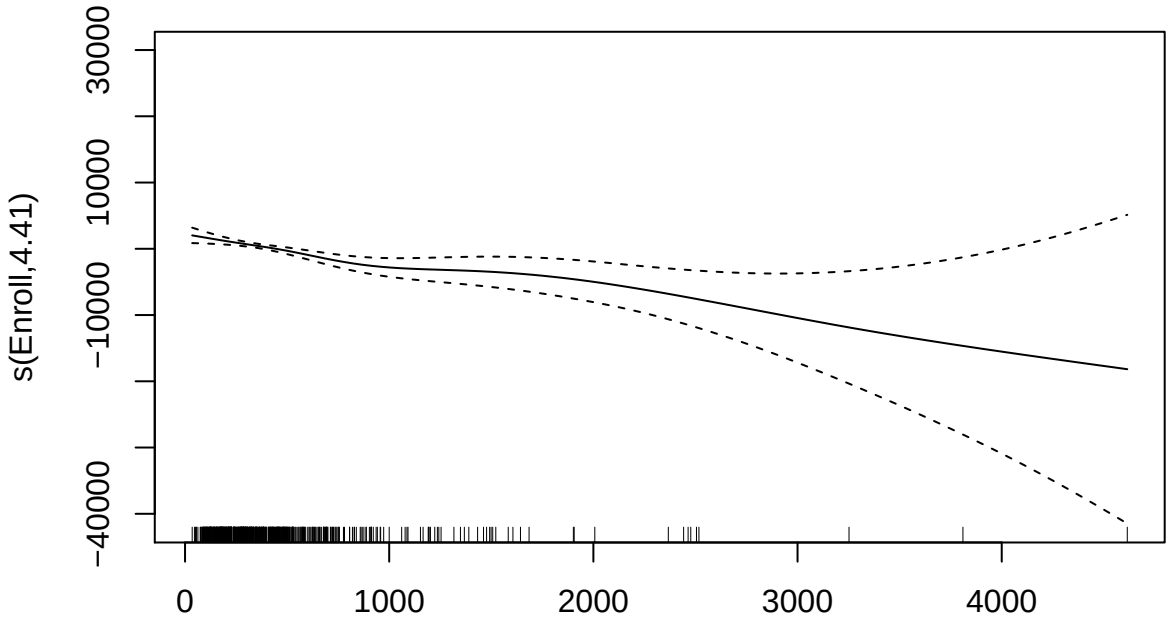
```
plot(gam.m2)
```

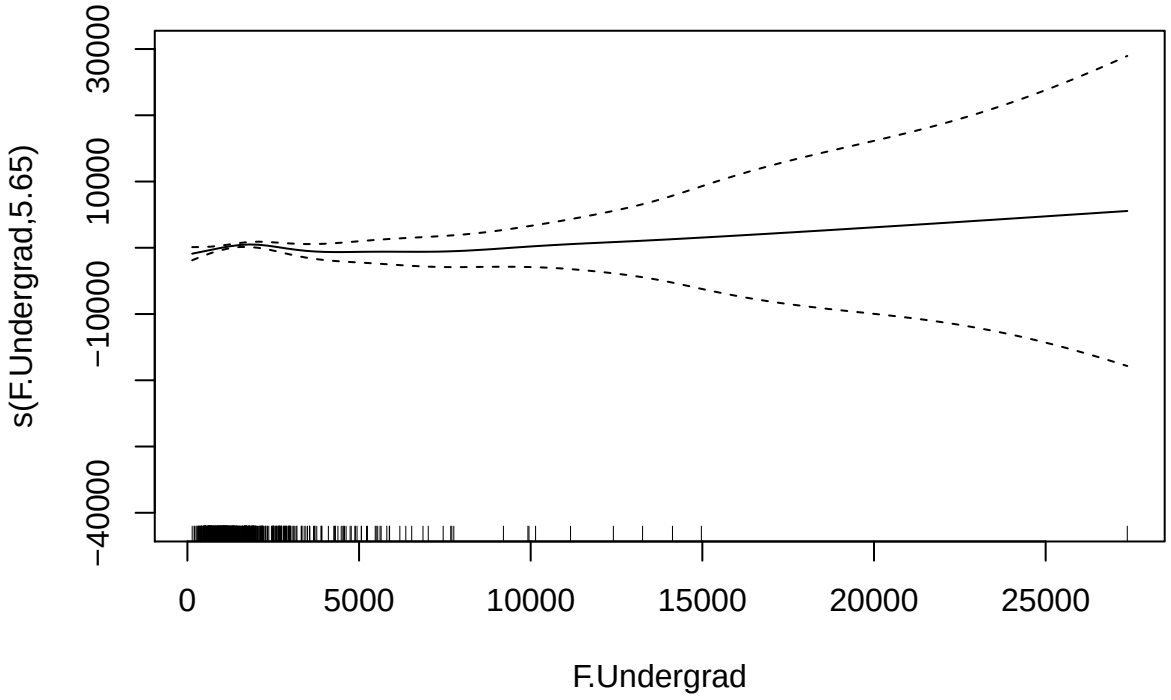


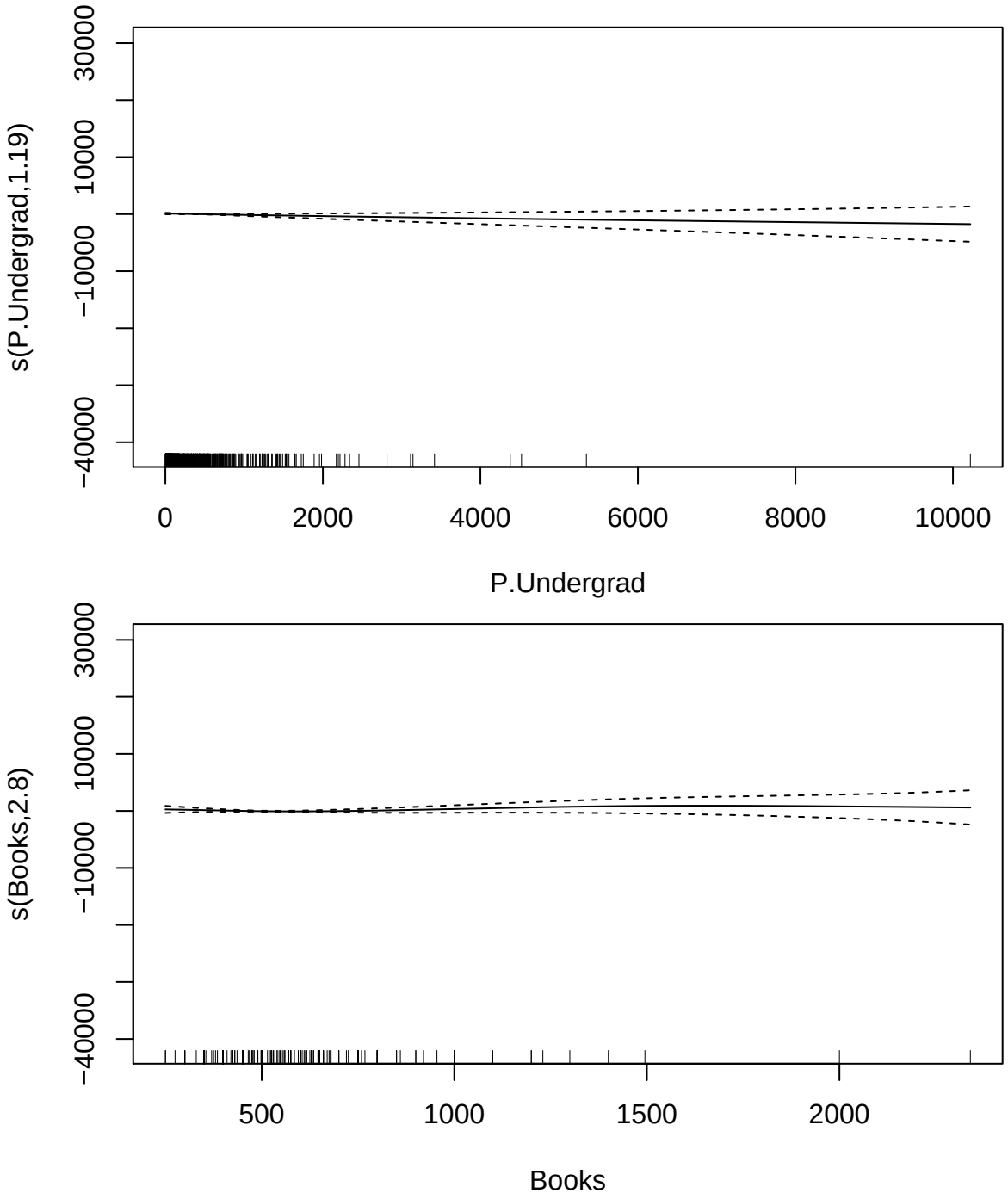

```
plot(gam.m3)
```

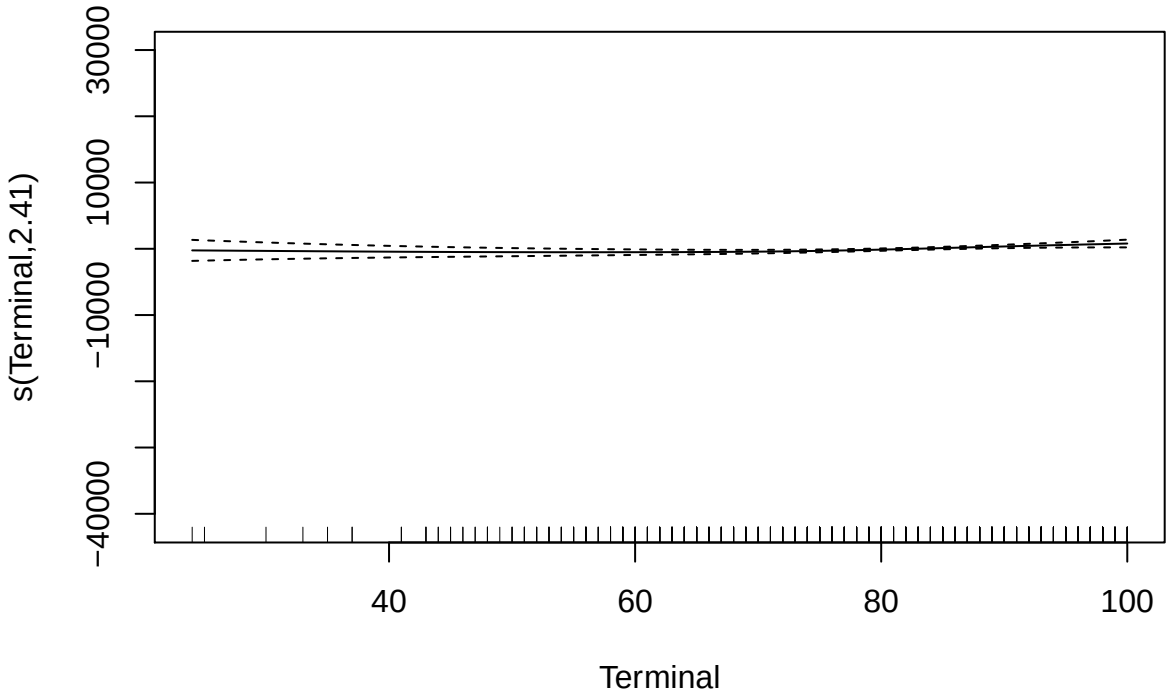

s(Apps,1)
30000
10000
−10000
−40000
0
5000
10000
15000
20000
Apps


s(Accept,4.39)
30000
10000
−10000
−40000
0
2000
4000
6000
8000
10000
12000
Accept

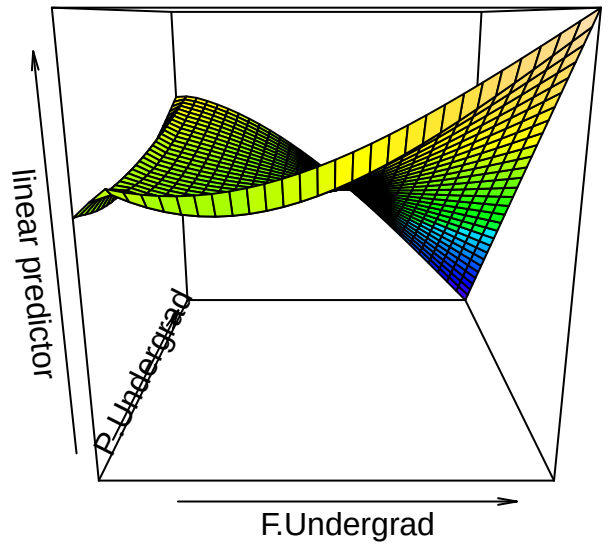

s(Enroll,4.41)
30000
10000
−10000
−40000
0
1000
2000
3000
4000
Enroll


s(F.Undergrad,5.65)
30000
10000
−10000
−40000
0
5000
10000
15000
20000
25000
F.Undergrad

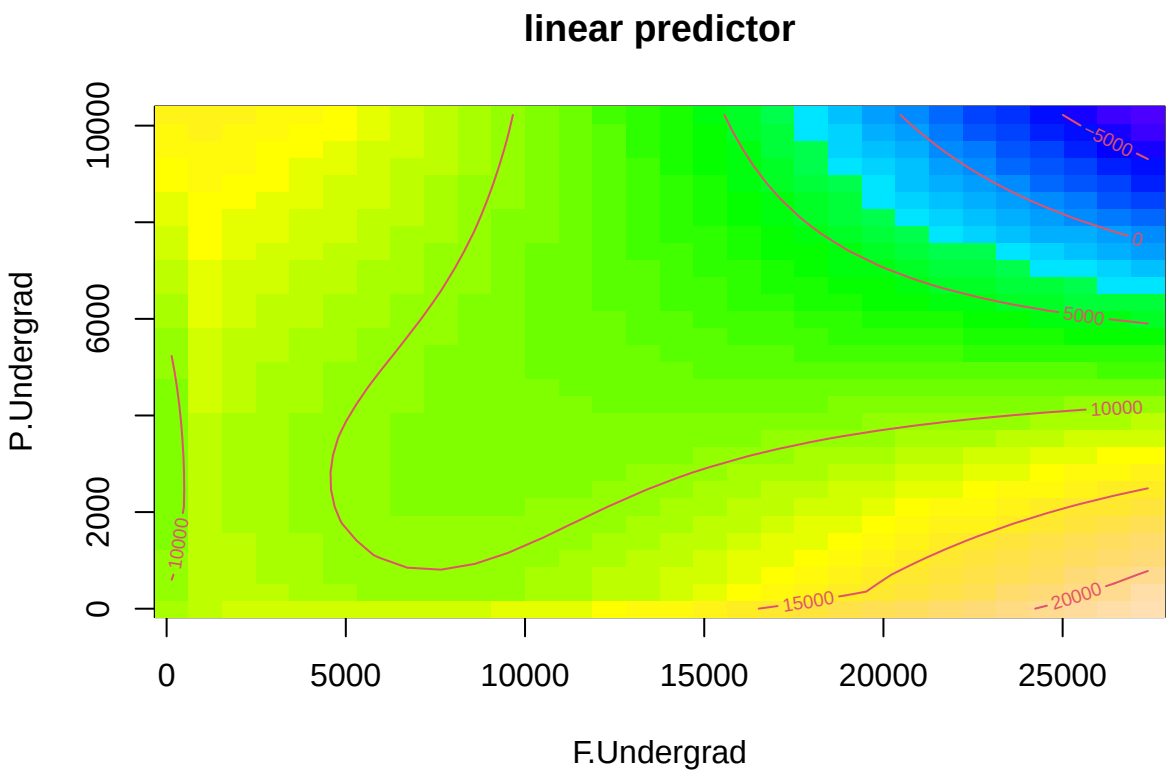

s(P.Undergrad,1.19)
30000
10000
−10000
−40000
0
2000
4000
6000
8000
10000
P.Undergrad
s(Books,2.8)
30000
10000
−10000
−40000
500
1000
1500
2000
Books




```
vis.gam(gam.m4, view=c("F.Undergrad","P.Undergrad"), color = "topo")
```



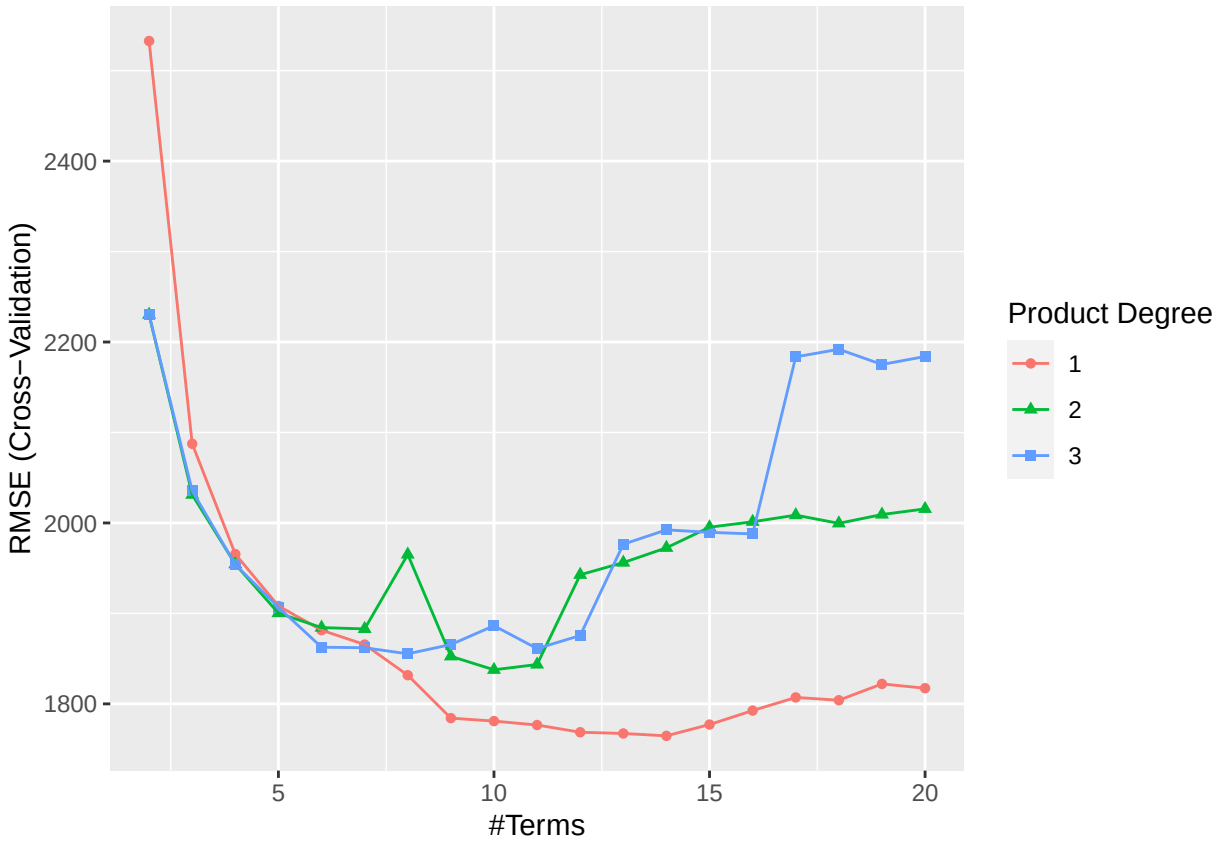

```
vis.gam(gam.m4, view=c("F.Undergrad","P.Undergrad"), color = "topo", plot.type = "contour")
```



Description: I have fitted multiple GAM models using all the predictors. From running the ANOVA test and checking the p-values of the result output, I could see that the third GAM model(gam.m3) shows the best fit. A plot of the second GAM model(gam.m2) showed the plot of the s function of the 'Terminal' predictor. The plot was neither noticeably increasing/decreasing until around a 'Terminal' value of 80, and then showed an increasing trend as 'Terminal' value increased. The credible interval spread out towards the ends of the plot. I have also plotted all the s functions of the third GAM model. The plot of the s function of 'Accept' showed a slightly increasing trend and the credible interval spread out quite a lot as the values of the predictor increased. However, the plot of the s function of 'Enroll' showed a slgitly decreasing trend and the credible interval spread out quite a lot as well as the values of the predictor increased. For all the other plots of the s functions of each of their corresponding predictors, the plots did not noticeably increase or decrease. I have also plotted the te function of 'F.Undergrad' and 'P.Undergrad' predictors, and the different colors indicate the magnitude of the functions.

(d)Train a MARS model using all the predictors

```
ctrl1 <- trainControl(method = "cv", number = 10)

mars_grid <-expand.grid(degree = 1:3,
                        nprune = 2:20)
set.seed(1)
mars.fit <- train(x, y,
                  method = "earth",
                  tuneGrid = mars_grid,
                  trControl = ctrl1)
ggplot(mars.fit)
```




```
mars.fit$bestTune
```

```
##    nprune degree
## 13     14      1
```

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

```
##          (Intercept)     h(Expend-15365)  h(4450-Room.Board)     h(Grad.Rate-97)
##        10661.2171202          -0.7327991          -1.2690141        -240.6920324
##     h(97-Grad.Rate) h(F.Undergrad-1355) h(1355-F.Undergrad)   h(22-perc.alumni)
##         -24.5321278          -0.3593198          -1.6238125         -78.4307664
##        h(Apps-3712)    h(1300-Personal)       h(913-Enroll)      h(2193-Accept)
##           6.9809000           1.0421976           5.3654412          -1.9460457
##      h(Expend-6881)        h(Apps-3877)
##           0.7300921          -6.6266893
```

For the final model, the nprune value is 14 and degree is 1. The final model is Intercept -240.692034 * h(Grad.Rate-97) -24.5321278 * h(97- Grad.Rate) -0.3593198 * h(F.Undergrad-1355) -1.6238125 * h(1355-F.Undergrad). The hinge functions are h(Grad.Rate-97), h(97- Grad.Rate), and h(F.Undergrad-1355), h(1355-F.Undergrad).

(d)Present the partial dependence plots

```
p1<- pdp::partial(mars.fit, pred.var = c("Grad.Rate"), grid.resolution = 10) %>% autoplot()
p2 <- pdp::partial(mars.fit, pred.var = c("F.Undergrad"), grid.resolution = 10) %>% autoplot()

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

```
## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.

## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.

## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.

## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.
```