P8106_HW2

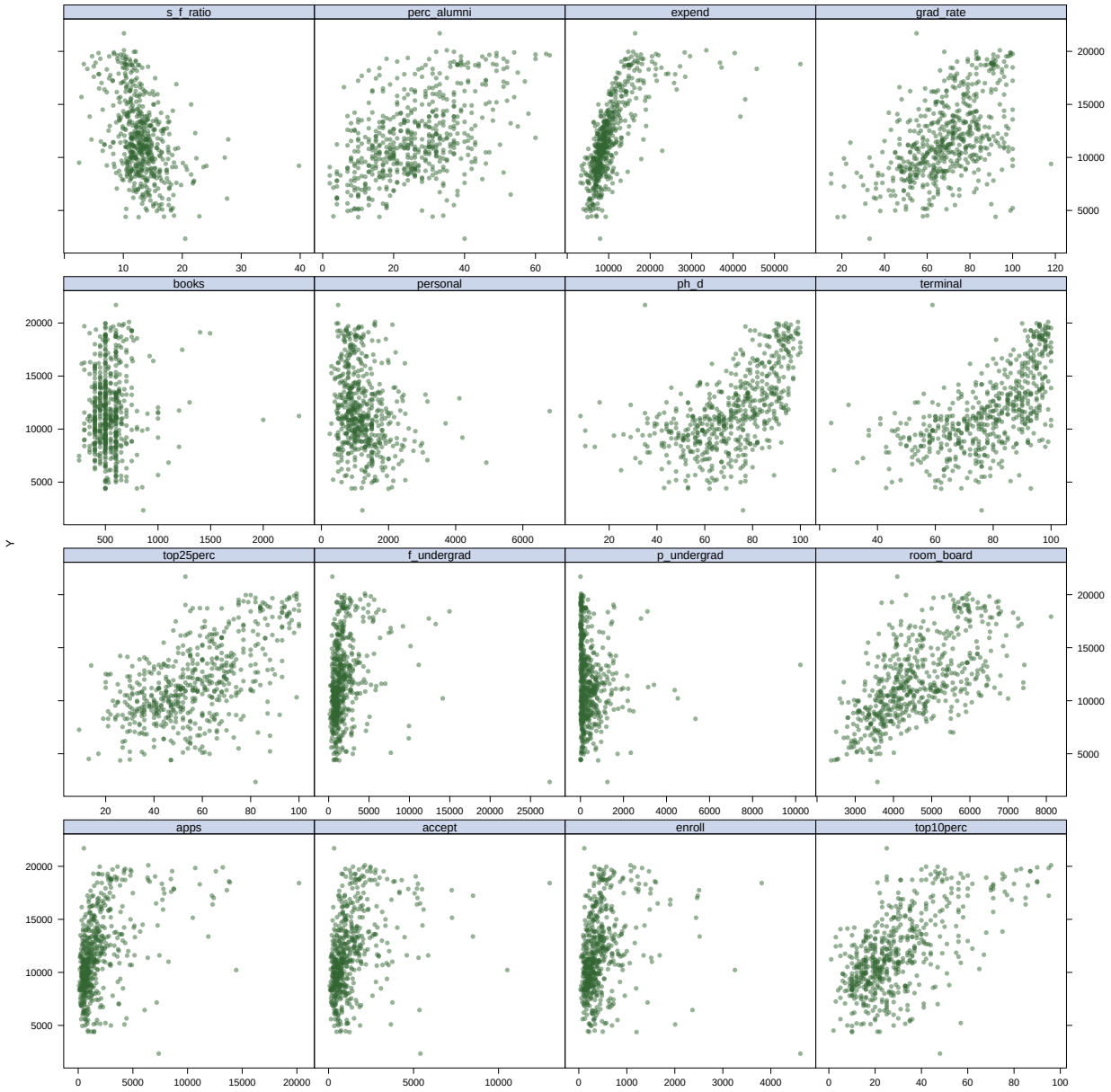
Ziyi Zhao

3/19/2020

```
college <- read_csv("./College.csv") %>% janitor::clean_names()

train_college <- filter(college,college!="Columbia University")

## matrix of predictors
x <- model.matrix(outstate~.,train_college)[,c(565:580)]

## vector of response
y <- train_college$outstate
```

(a) Create scatter plots of response vs. predictors.

```
theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(0.2,0.4,0.2,0.5)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(0.8,0.1,0.1,0.1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(0.0,0.2,0.6,0.2)
trellis.par.set(theme1)
featurePlot(x,y,plot="scatter",labels = c("","Y"),
            type=c("p"),layout=c(4,4))
```

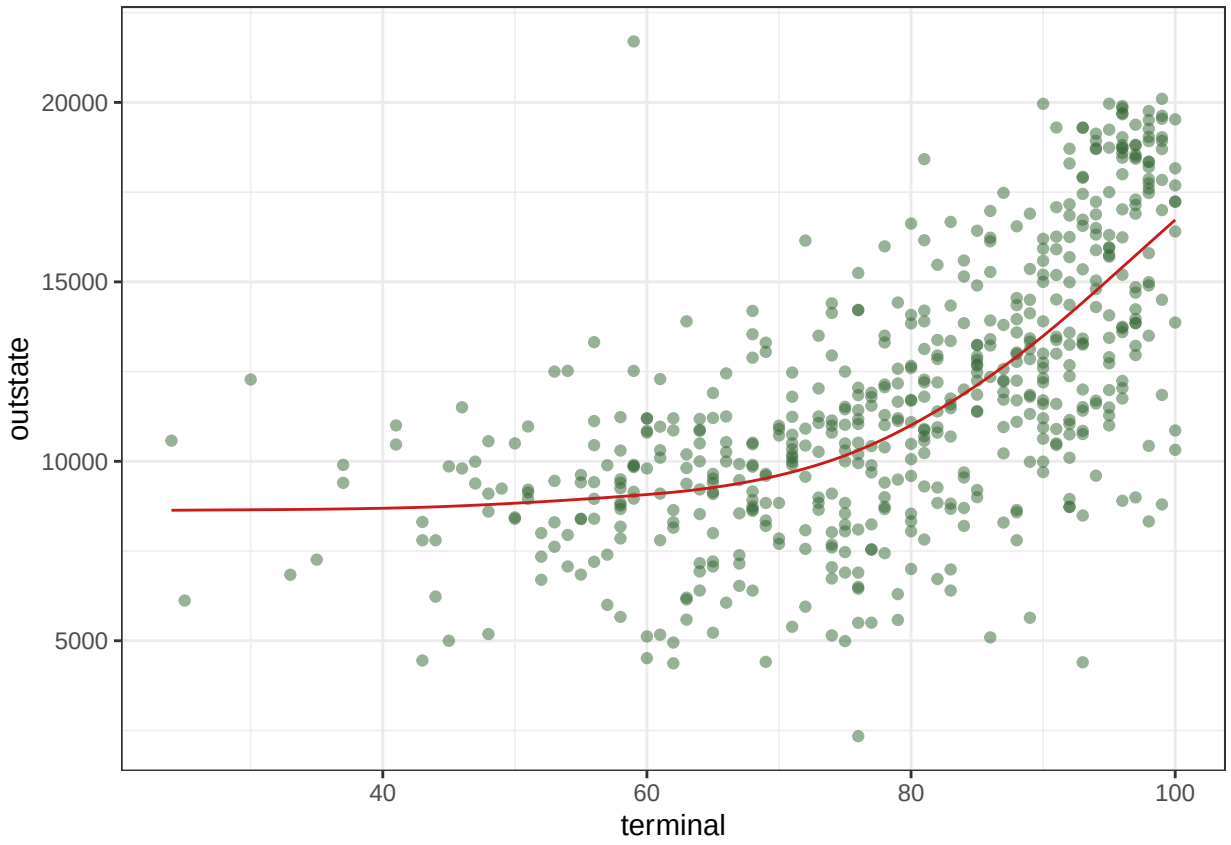


(b) Fit a smoothing spline model using Terminal as the only predictor of Outstate for a range of degrees of freedom, as well as the degree of freedom obtained by generalized cross validation, and plot the resulting fits. Describe the results obtained.

```
fit.ss <- smooth.spline(train_college$terminal,train_college$outstate)
fit.ss$df
```

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

```
terminallims <- range(train_college$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=train_college,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()
```



According to the output, the degreee of freedom obtained by generalized cross-validation is 4.4686294.

The solution of smooth function g(x) is a natural cubic spline with knots at every unique value of variable terminal ranged from 24 to 100.

From the plot, the scatter plot shown in green points is the train dataset of college with terminal as x-axis and outstate as y-axis. The red fitted curve is generated by predicted dataset of outstate by every unique value of terminal. The predicted curve fits the data smoothly.

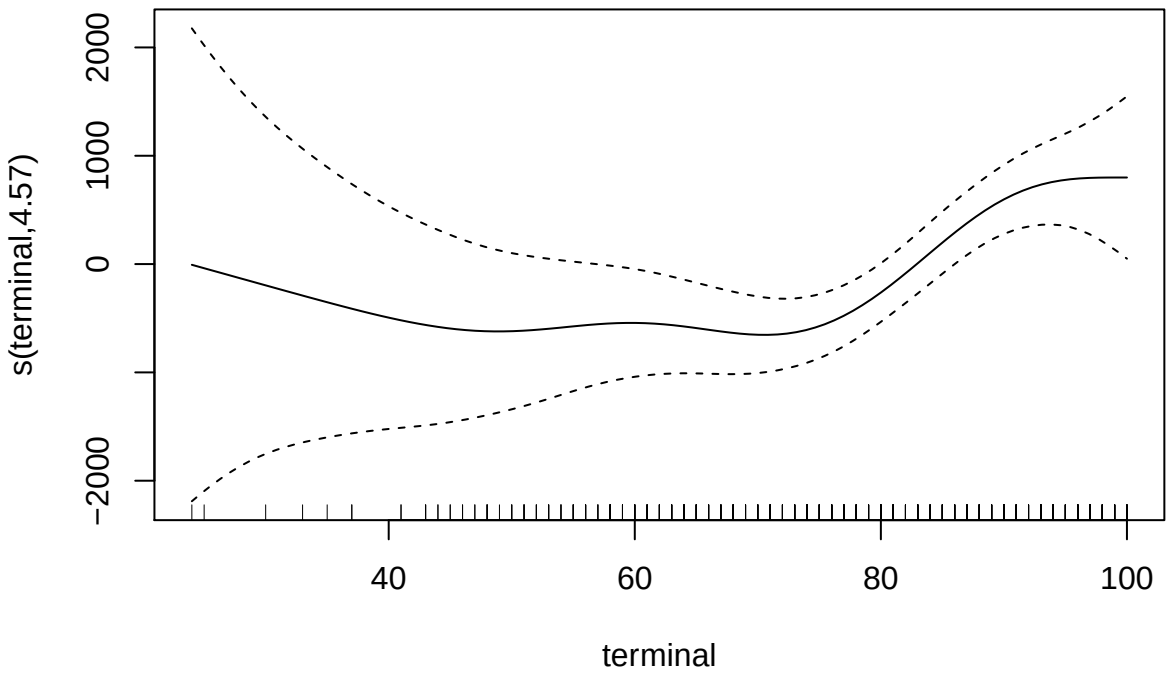
(c) Fit a generalized additive model (GAM) using all the predictors. Plot the results and explain your findings.

Use gam():

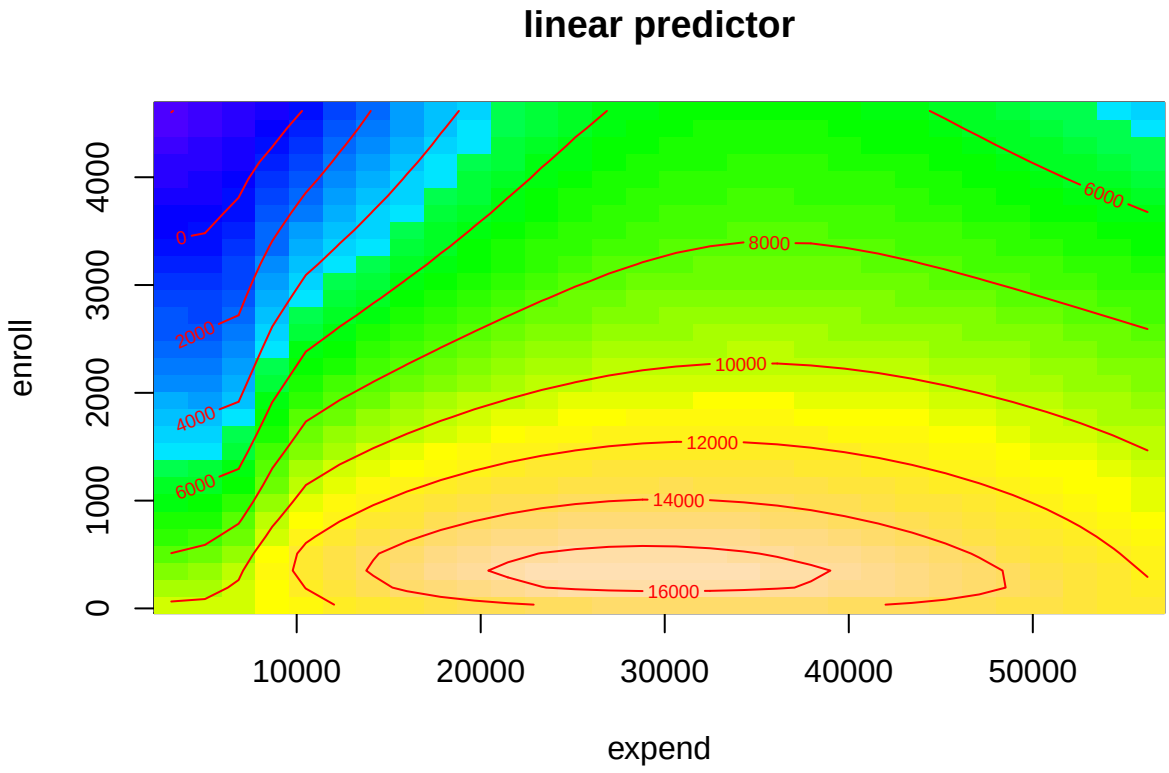
```
gam.m1 <- gam(outstate~apps+accept+enroll+top10perc+top25perc+
                f_undergrad+p_undergrad+room_board+books+personal+ph_d+
                terminal+s_f_ratio+perc_alumni+expend+grad_rate,
              data=train_college)
gam.m2 <- gam(outstate~apps+accept+enroll+top10perc+top25perc+
                f_undergrad+p_undergrad+room_board+books+personal+ph_d+
                s(terminal)+s_f_ratio+perc_alumni+expend+grad_rate,
              data=train_college)
gam.m3 <- gam(outstate~apps+accept+ph_d+top10perc+top25perc+
                f_undergrad+p_undergrad+books+personal+
                s(terminal)+s_f_ratio+te(expend,enroll)+
                perc_alumni+room_board+grad_rate,
              data=train_college)

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

```
## Analysis of Deviance Table
##
## Model 1: outstate ~ apps + accept + enroll + top10perc + top25perc + f_undergrad +
##     p_undergrad + room_board + books + personal + ph_d + 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 + ph_d + s(terminal) +
##     s_f_ratio + perc_alumni + expend + grad_rate
## Model 3: outstate ~ apps + accept + ph_d + top10perc + top25perc + f_undergrad +
##     p_undergrad + books + personal + s(terminal) + s_f_ratio +
##     te(expend, enroll) + perc_alumni + room_board + grad_rate
##   Resid. Df Resid. Dev     Df  Deviance       F    Pr(>F)
## 1    547.00 2092185295
## 2    542.37 2026858216 4.6295  65327078  4.7398 0.0004541 ***
## 3    532.66 1591448554 9.7101 435409662 15.0619 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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




```
vis.gam(gam.m3,view = c("expend","enroll"),
        plot.type = "contour",color="topo")
```



Use caret:

```
ctrl1 <- trainControl(method="cv",number=5)

set.seed(7)
gam.fit <- train(x,y,
                 method="gam",
                 tuneGrid = data.frame(method = "GCV.Cp",select=c(TRUE,FALSE)),
                 trControl = ctrl1)

gam.fit$bestTune
```

```
##   select method
## 1  FALSE GCV.Cp
```

```
gam.fit$results
```

```
##   method select     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1 GCV.Cp  FALSE 1816.709 0.7654877 1384.619 204.1885 0.04437825 107.4501
## 2 GCV.Cp   TRUE 1905.812 0.7476713 1415.683 301.6884 0.06002304 101.9745
```

```
gam.fit$finalModel
```

```
##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(perc_alumni) + s(terminal) + s(top10perc) + s(ph_d) +
##     s(grad_rate) + s(books) + s(top25perc) + s(s_f_ratio) + s(personal) +
##     s(p_undergrad) + s(enroll) + s(room_board) + s(accept) +
##     s(f_undergrad) + s(apps) + s(expend)
##
## Estimated degrees of freedom:
## 1.90 5.14 3.64 6.32 4.27 2.35 1.00
## 4.33 1.00 1.00 1.00 2.13 3.58 6.28
## 4.59 6.45  total = 55.98
##
## GCV score: 2761951
```

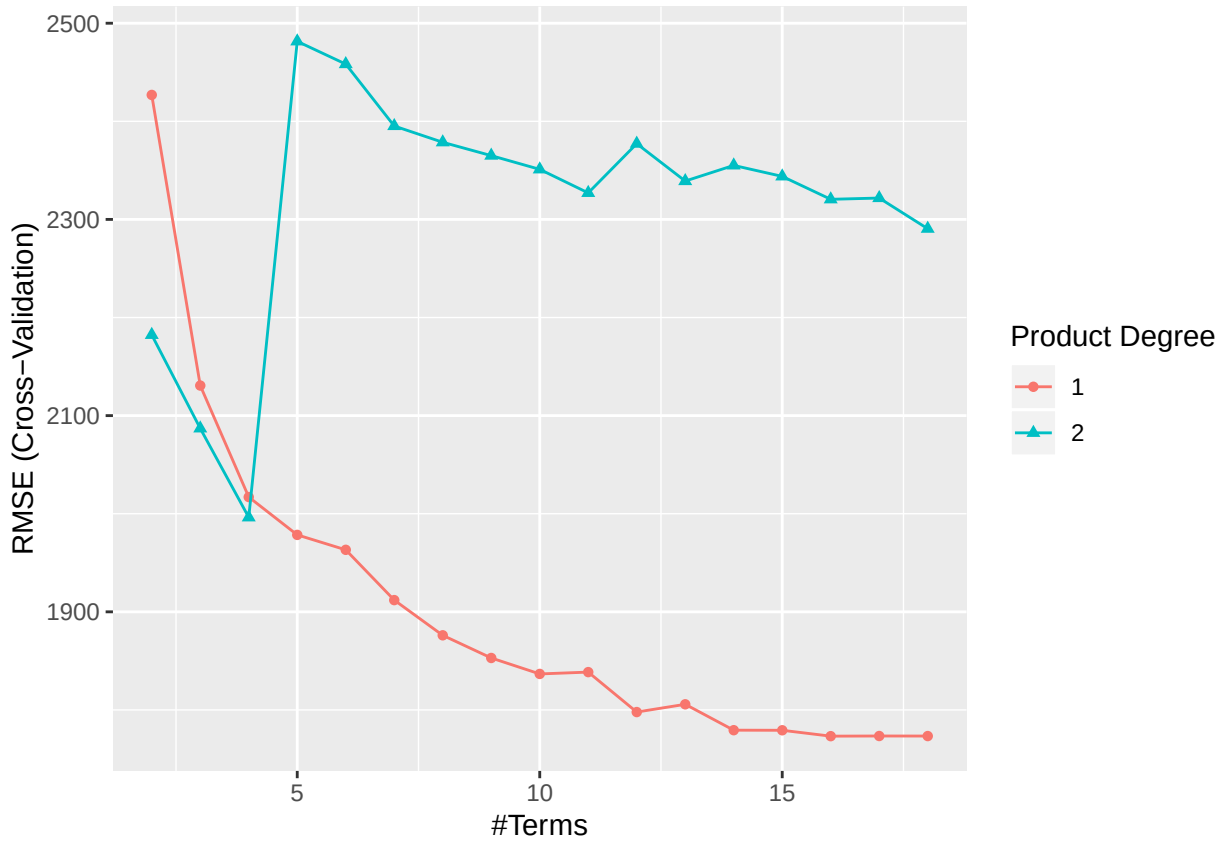
I separately use gam() and caret to build models to estimate the relationship between the outstate and predictors.

From the results of using caret, we can see the output of bestTune showed that selecting "False" is better than selecting "True". The results of gam.fit also showed that rmse of selecting "false" is smaller than that of selecting "true". I think it may be caused by loss of significant amount of flexibility in mgcv.From the final model of gam.fit, it added smooth function to every variable. Both df and GCV score are very large.

In anova test of three gam(), the model 2 with smoothed terminal has df = 4.6295, which is approximate to the df obtained by using smoothe spline model on terminal only. It has smaller deviance than model 3. The result of caret may indicate there are some potential tensor interaction between predictors.

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.

```
mars_grid <- expand.grid(degree=1:2,
                         nprune=2:18)

set.seed(7)
mars.fit <- train(x,y,
                  method = "earth",
                  tuneGrid = mars_grid,
                  trControl = ctrl1)

ggplot(mars.fit)
```




```
mars.fit$bestTune
```

```
##    nprune degree
## 15     16      1
```

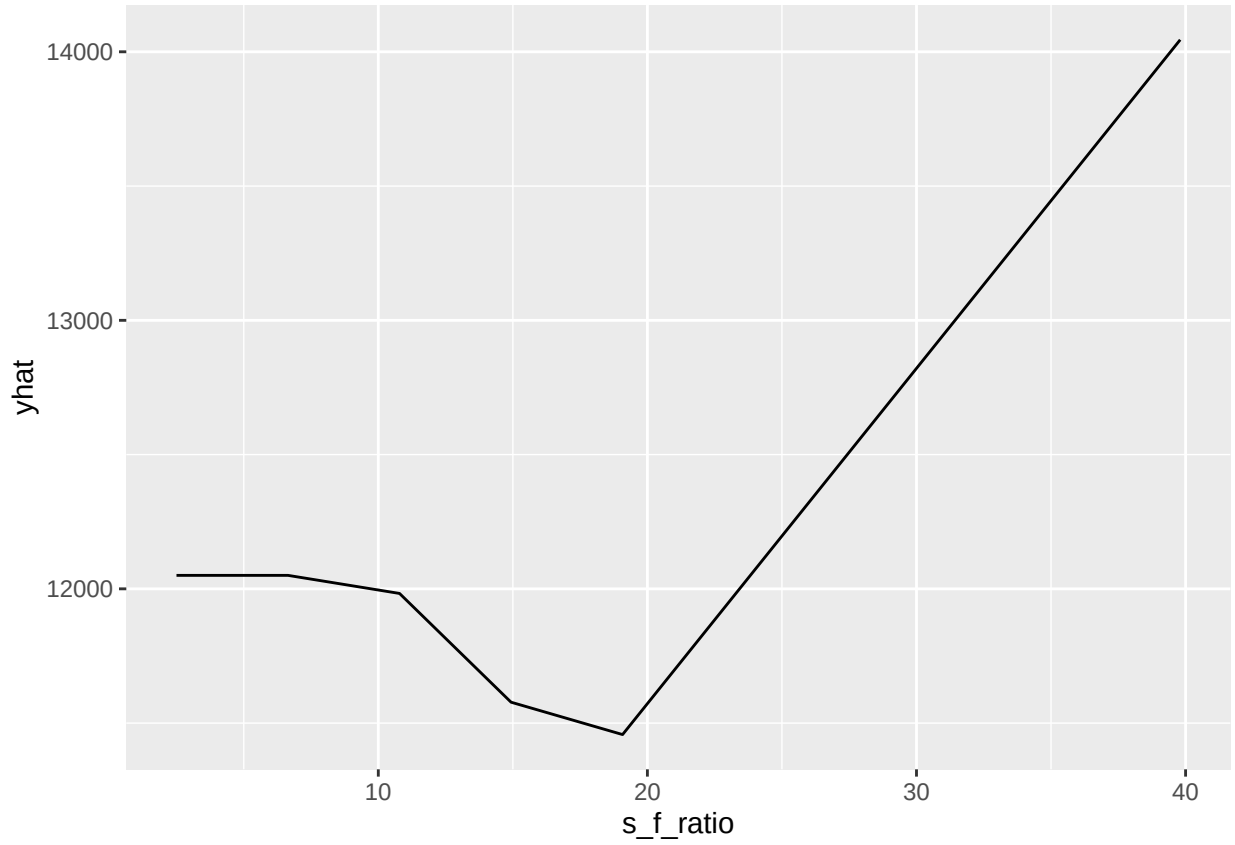
```
coef(mars.fit$finalModel)
```

```
##          (Intercept)     h(expend-15365)  h(4450-room_board)
##       11157.3323100          -0.6964270          -1.2721516
##     h(grad_rate-97)     h(97-grad_rate) h(f_undergrad-1355)
##        -242.9852028         -24.1380627          -0.3567241
## h(1355-f_undergrad)   h(22-perc_alumni)        h(apps-3712)
##          -1.7564383         -77.0359905           7.0815293
##    h(1300-personal)       h(913-enroll)      h(2193-accept)
##           1.0492662           5.2944664          -1.9951560
##      h(expend-6881)        h(apps-3877)   h(s_f_ratio-10.1)
##           0.6896165          -6.7393719         -97.7224105
##   h(s_f_ratio-17.8)
##         222.5913868
```

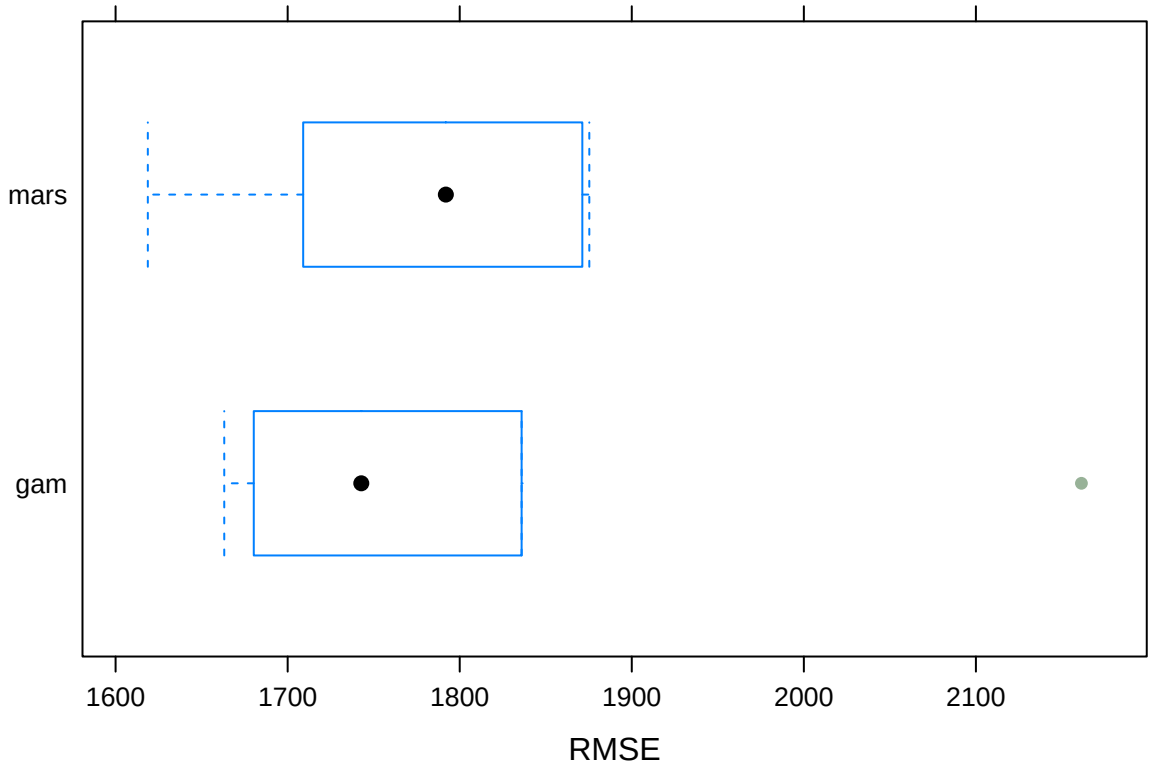
The final model using MARS:

f(x) = 11157.33 - 0.6964h(expend-15365) - 1.2722h(4450-room_board) - 242.9852h(grad_rate-97) - 24.1381h(97-grad_rate) - 0.3567h(f_undergrad-1355) - 1.7564h(1355-f_undergrad) - 77.0360h(22-perc_alumni) + 7.0815h(apps-3712) + 1.0493h(1300-personal) + 5.2945h(913-enroll) - 1.9952h(2193-accept) + 0.6896h(expend-6881) - 6.7394h(apps-3877) - 97.7224h(s_f_ratio-10.1) + 222.5914h(s_f_ratio-17.8)

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

p1
```




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



According to the boxplot, the GAM (using caret) has smaller RMSE than the MARS does.

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

```
columbia <-  filter(college,college=="Columbia University")

x_col <- select(columbia,-c(college,outstate))

pred_gam <- predict(gam.fit,newdata = x_col) ## using gam.fit by caret
pred_gam
```

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

```
pred_gam_m2 <- predict(gam.m2,newdata = x_col) ## using gam.m2
pred_gam_m2
```

```
##        1
## 19406.71
```

```
pred_gam_m3 <- predict(gam.m3, newdata = x_col) ## using gam.m3
pred_gam_m3
```

```
##        1
## 19433.87
```

```
pred_mars <- predict(mars.fit,newdata = x_col)
pred_mars
```

```
##            y
## [1,] 18520.5
```

Using the GAM, the predicted out-of-state tuition is 1.7728506×10^4 dollars.

Using the MARS, the predicted out-of-state tuition is 1.8520501×10^4 dollars.