

Homework 2

Yiying Wu (yw3996)

R packages

```
library(tidyverse)
library(caret)
library(tidymodels)
library(splines)
```

dataset

```
dat<-read_csv("./data/College.csv")
dat <- na.omit(dat)
```

Partition the dataset into two parts: training data (80%) and test data (20%).

```
set.seed(1)
data_split <- initial_split(dat, prop = 0.80)

# Extract the training and test data
training_data <- training(data_split)
x_train <- training_data %>% select(-College, -Outstate)
y_train <- training_data$Outstate

testing_data <- testing(data_split)
x_test <- testing_data %>% select(-College, -Outstate)
y_test <- testing_data$Outstate

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

Outcome variable: Outstate

(a)

Fit smoothing spline models to predict out-of-state tuition (Outstate) using the percentage of alumni who donate (perc.alumni) as the only predictor, across a range of degrees of freedom. Plot the model fits for each degree of freedom. Describe the observed patterns that emerge with varying degrees of freedom. Select an appropriate degree of freedom for the model and plot this optimal fit. Explain the criteria you used to determine the best choice of degree of freedom.

```
range(dat$perc.alumni)
```

```
## [1]  2 64
```

```
perc.alumni.grid <- seq(from = 0, to = 65, by = 1)
```

```
# plot the scatter plot for the training data
p <- ggplot(data = training_data, aes(x = perc.alumni, y = Outstate)) +
  geom_point(color = rgb(.2, .4, .2, .5))
```

For degree of freedom = 3

```
fit.ss3 <- smooth.spline(training_data$perc.alumni, training_data$Outstate,df=3)

pred.ss3 <- predict(fit.ss3,
  x = perc.alumni.grid)

pred.ss.df3 <- data.frame(pred = pred.ss3$y,
  perc.alumni = perc.alumni.grid)
```

For degree of freedom = 5

```
fit.ss5 <- smooth.spline(training_data$perc.alumni, training_data$Outstate,df=5)

pred.ss5 <- predict(fit.ss5,
  x = perc.alumni.grid)

pred.ss.df5 <- data.frame(pred = pred.ss5$y,
  perc.alumni = perc.alumni.grid)
```

For degree of freedom = 8

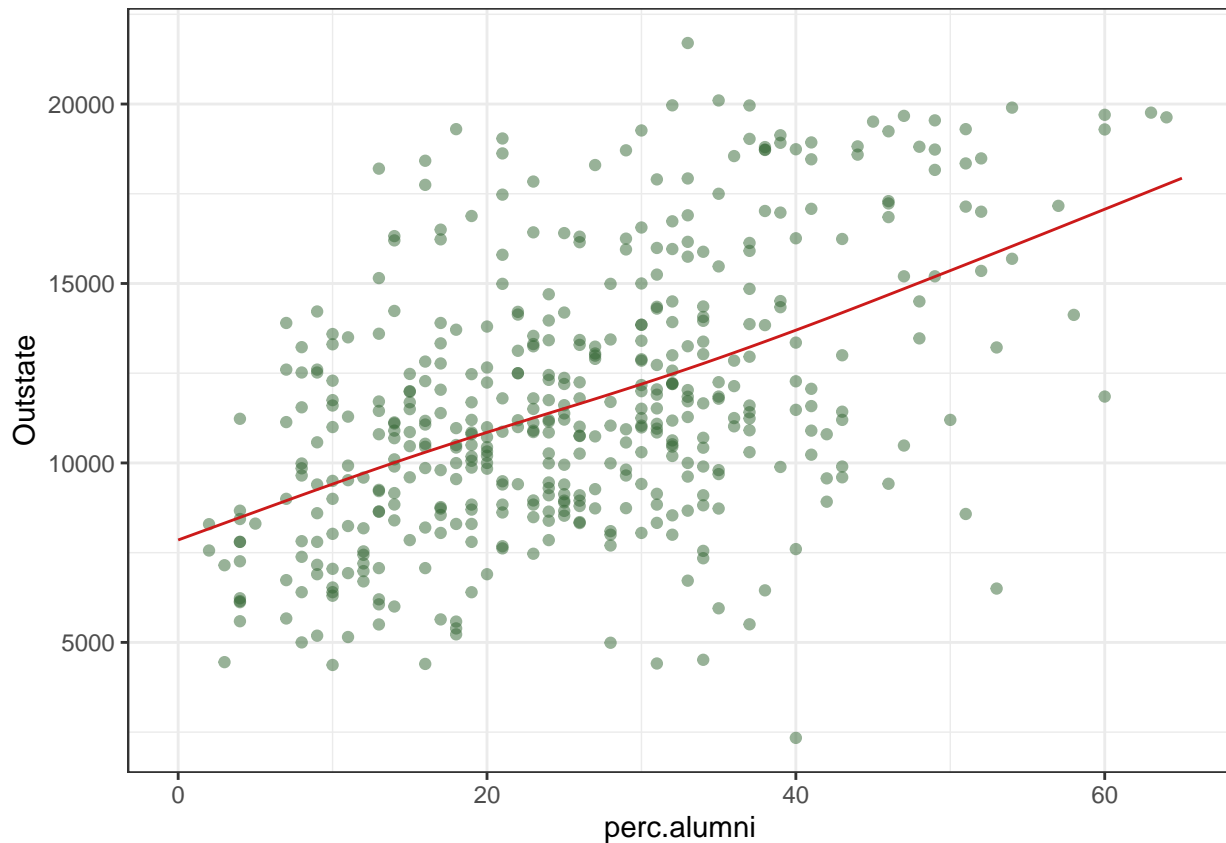
```
fit.ss8 <- smooth.spline(training_data$perc.alumni, training_data$Outstate,df=8)

pred.ss8 <- predict(fit.ss8,
  x = perc.alumni.grid)

pred.ss.df8 <- data.frame(pred = pred.ss8$y,
  perc.alumni = perc.alumni.grid)
```

Plot the model fits for df=3

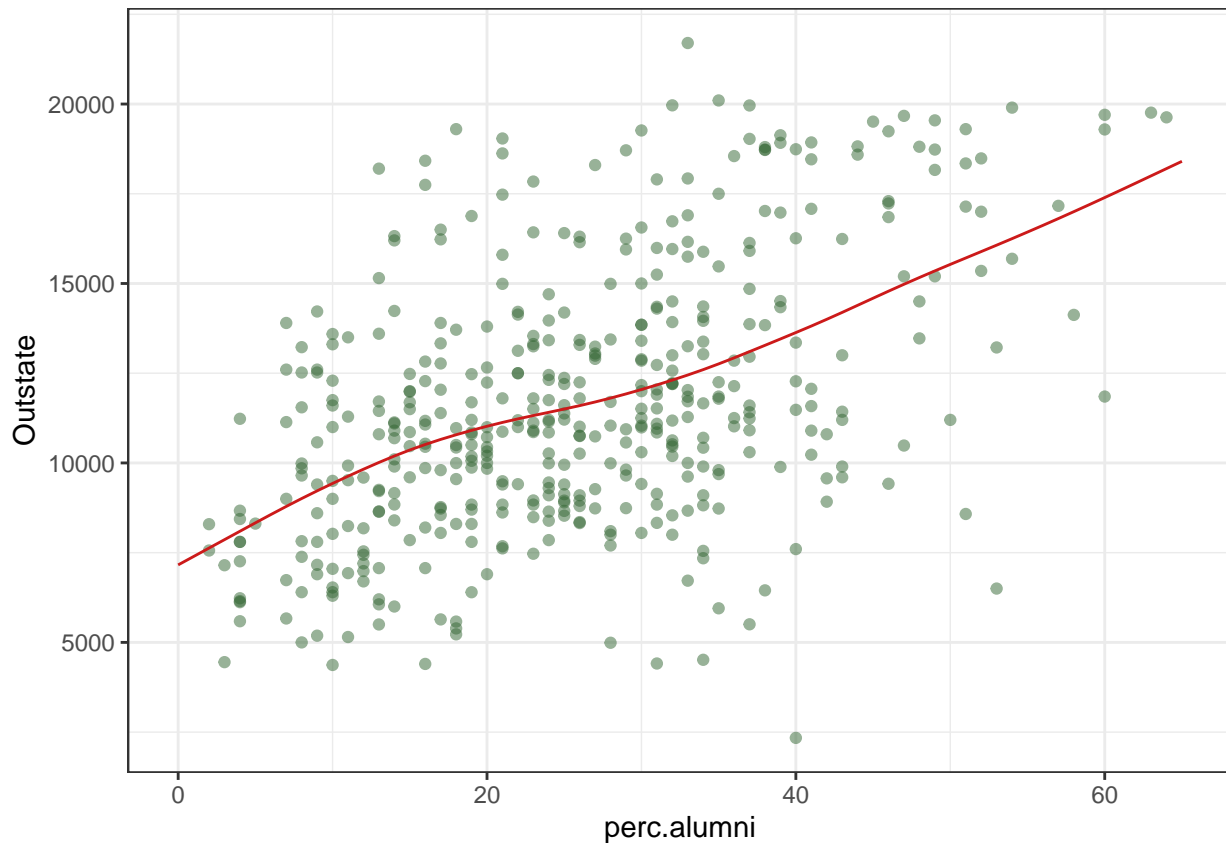
```
p +
  geom_line(aes(x = perc.alumni, y = pred), data = pred.ss.df3,
    color = rgb(.8, .1, .1, 1)) + theme_bw()
```



Description From the plot, there seems to be a positive correlation between the percentage of alumni donors and out-of-state tuition costs. The relationship seems almost linear.

Plot the model fits for $df=5$

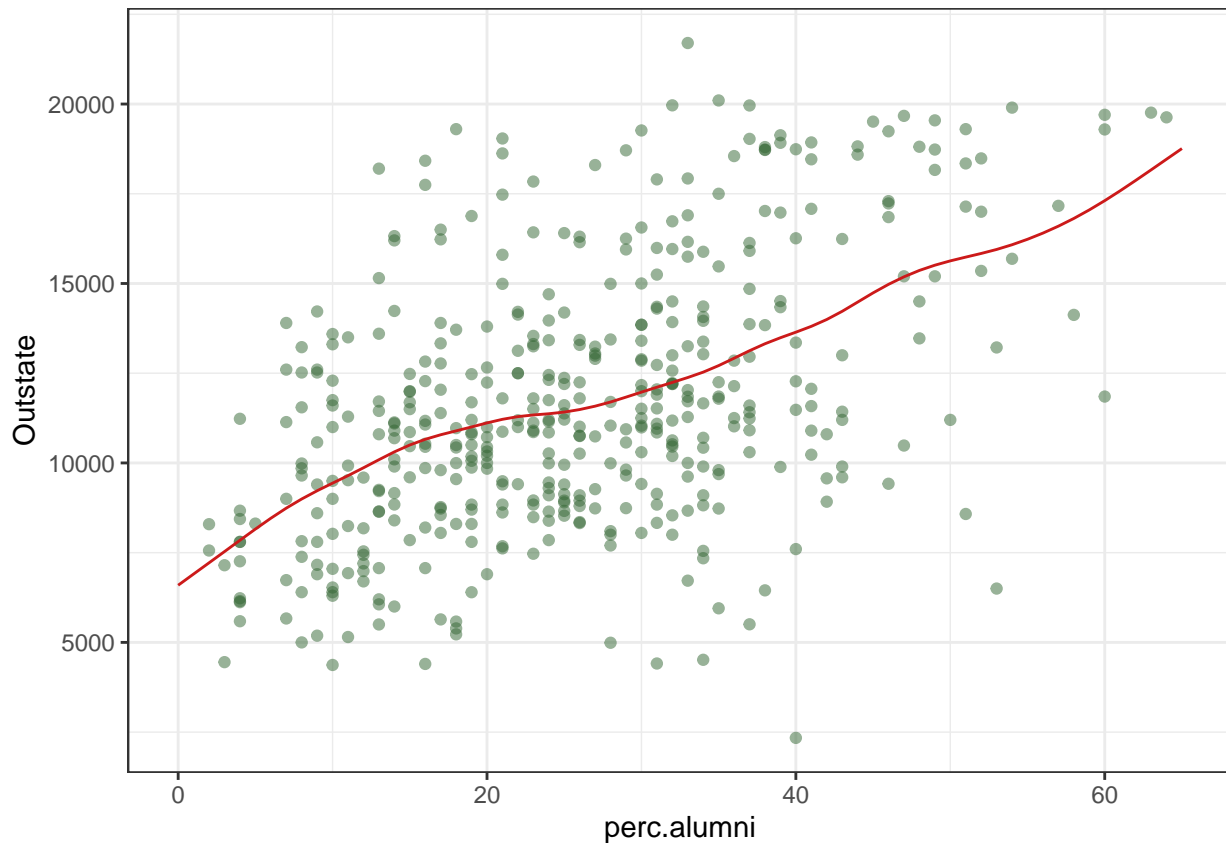
```
p +  
geom_line(aes(x = perc.alumni, y = pred), data = pred.ss.df5,  
          color = rgb(.8, .1, .1, 1)) + theme_bw()
```



Description In this plot, the fitted line seems to curve slightly upward from 5 to 30, suggesting a possible non-linear relationship. This might imply that the rate at which out-of-state tuition increases becomes greater at higher percentages of alumni donors within this range.

Plot the model fits for df=8

```
p +  
geom_line(aes(x = perc.alumni, y = pred), data = pred.ss.df8,  
          color = rgb(.8, .1, .1, 1)) + theme_bw()
```



Description This plot indicates a nonlinear pattern.

Select an appropriate degree of freedom for the model

```
# does generalized-cv to determine df automatically
fit.ss <- smooth.spline(training_data$perc.alumni, training_data$Outstate, df=3)

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

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

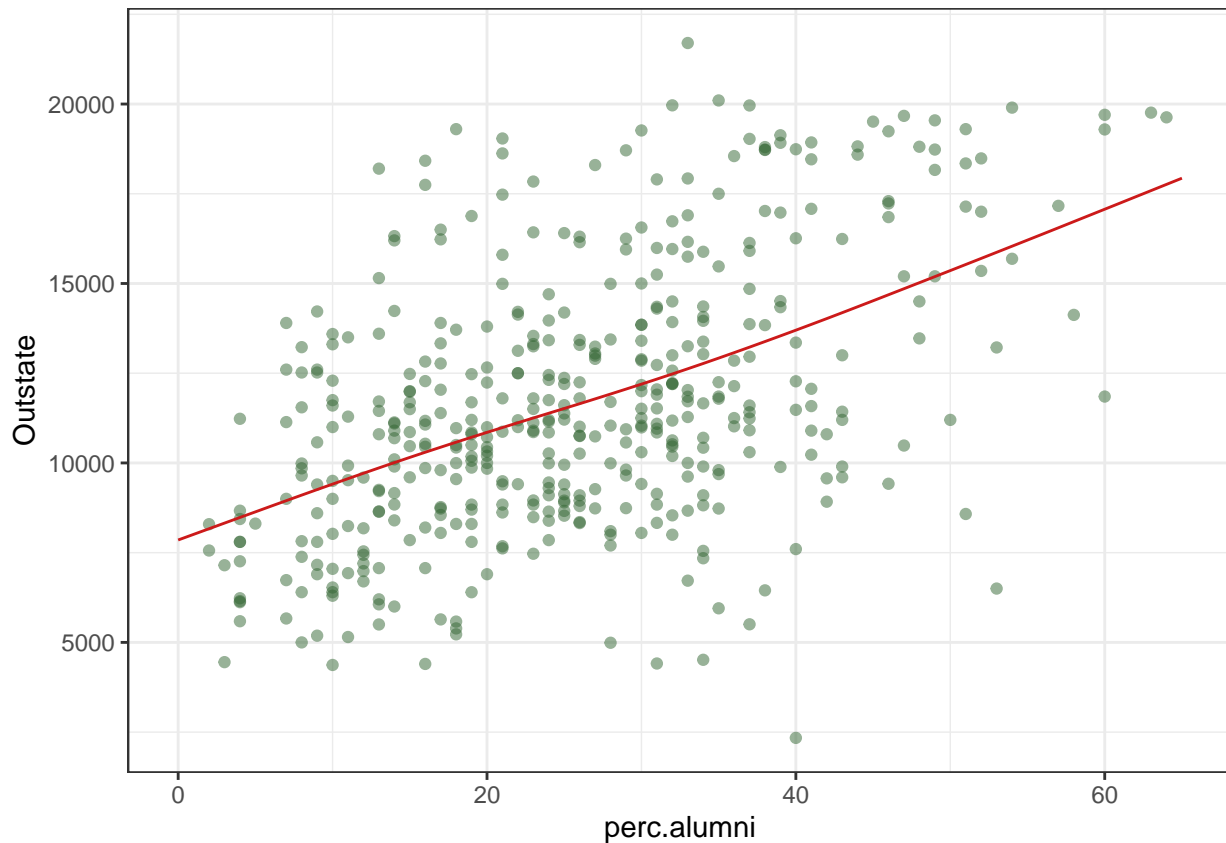
# print the selected df
fit.ss$df
```

```
## [1] 3.000318
```

Therefore, the appropriate degree of freedom for the model is 3.0003

plot this optimal fit

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



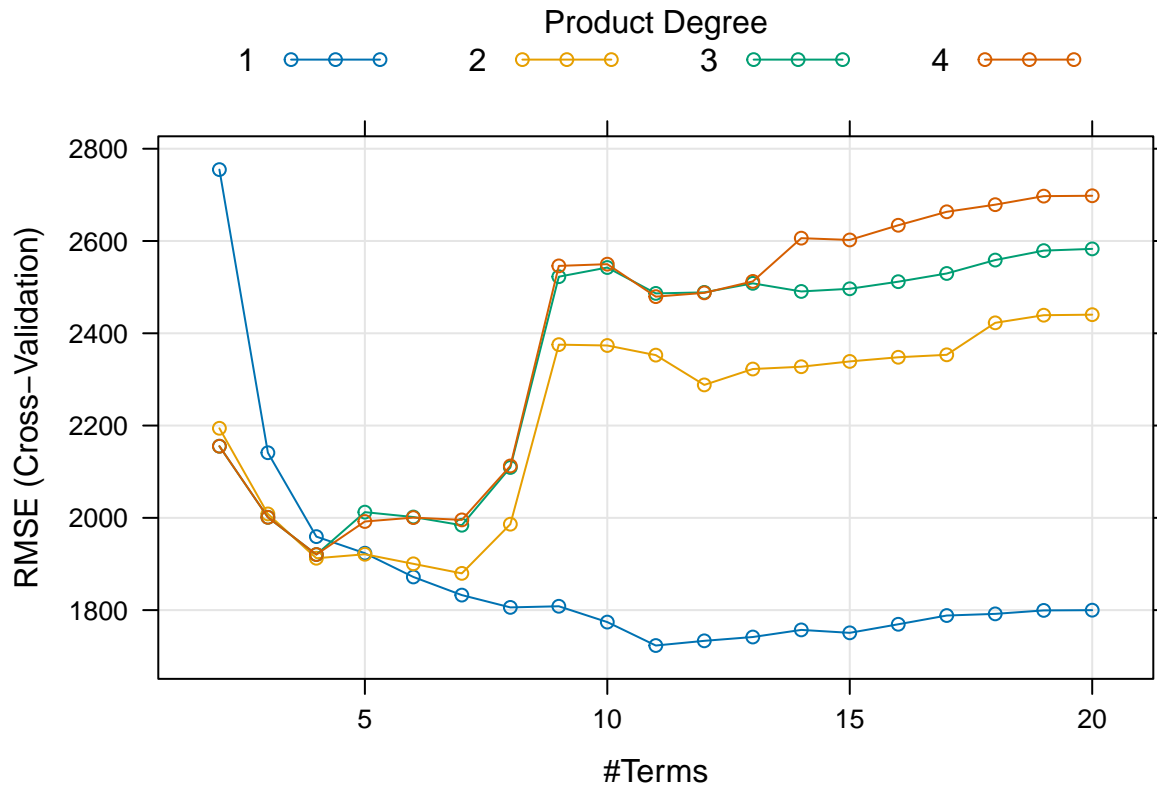
The criteria used to determine the best choice of degree of freedom is **generalized cross validation**.

(b)

Train a multivariate adaptive regression spline (MARS) model to predict the response variable. Report the regression function. Present the partial dependence plot of an arbitrary predictor in your model. Report the test error.

```
set.seed(1)
model.mars <- train(x = x_train,
                    y = y_train,
                    method = "earth", # earth is for mars
                    tuneGrid = expand.grid(degree = 1:4,
                                           nprune = 2:20),
                    trControl = ctrl)
# degree from 1~4 is sufficient
# nprune can be larger than the number of predictors, make it as large as possible

plot(model.mars)
```



```
# both number of terms and product degree are upper bounds
```

```
# best tune
```

```
model.mars$bestTune
```

```
##      nprune degree
```

```
## 10      11      1
```

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

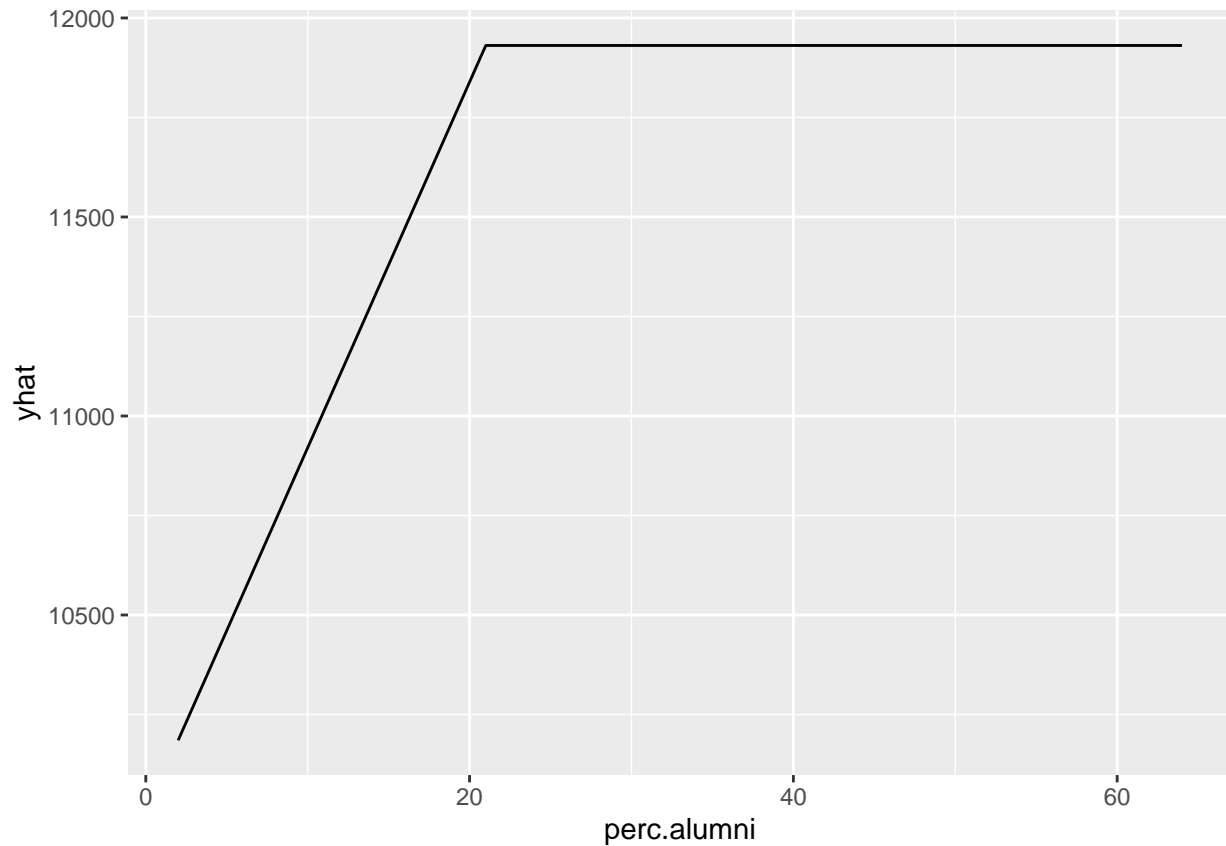
```
##      (Intercept)      h(Expend-15494)      h(81-Grad.Rate)      h(4138-Room.Board)
##      10252.3200044      -0.7594586      -29.5255903      -1.5537526
## h(F.Undergrad-1251) h(1251-F.Undergrad)      h(21-perc.alumni)      h(Apps-2694)
##      -0.3243249      -1.5277676      -91.8812340      0.3838076
##      h(942-Enroll)      h(Expend-6898)      h(2081-Accept)
##      5.1925945      0.7691120      -2.0869169
```

The regression function is

$$f(x) = 10252.32 - 0.76 \cdot h(\text{Expend} - 15494) - 29.53 \cdot h(81 - \text{Grad.Rate}) - 1.55 \cdot h(4138 - \text{Room.Board}) \\ - 0.32 \cdot h(F.Undergrad - 1251) - 1.53 \cdot h(1251 - F.Undergrad) - 91.88 \cdot h(21 - \text{perc.alumni}) \\ + 0.38 \cdot h(\text{Apps} - 2694) + 5.19 \cdot h(942 - \text{Enroll}) + 0.77 \cdot h(\text{Expend} - 6898) - 2.09 \cdot h(2081 - \text{Accept})$$

partial dependence plot of an arbitrary predictor (perc.alumni) in the model.

```
pdp::partial(model.mars, pred.var = c("perc.alumni"), grid.resolution = 200) %>% autoplot()
```



test error

```
mars.pred <- predict(model.mars, newdata = x_test)
test_error_mars <- mean((mars.pred - y_test)^2)
test_error_mars
```

```
## [1] 2990281
```

```
RMSE_mars <- sqrt(test_error_mars)
RMSE_mars
```

```
## [1] 1729.243
```

The RMSE of MARS model is 1729.24.

(c)

Construct a generalized additive model (GAM) to predict the response variable. Does your GAM model include all the predictors? For the nonlinear terms included in your model, generate plots to visualize these relationships and discuss your observations. Report the test error.


```
set.seed(1)
model.gam <- train(x = x_train,
                   y = y_train,
                   method = "gam",
                   trControl = ctrl)
```

```
model.gam$bestTune
```

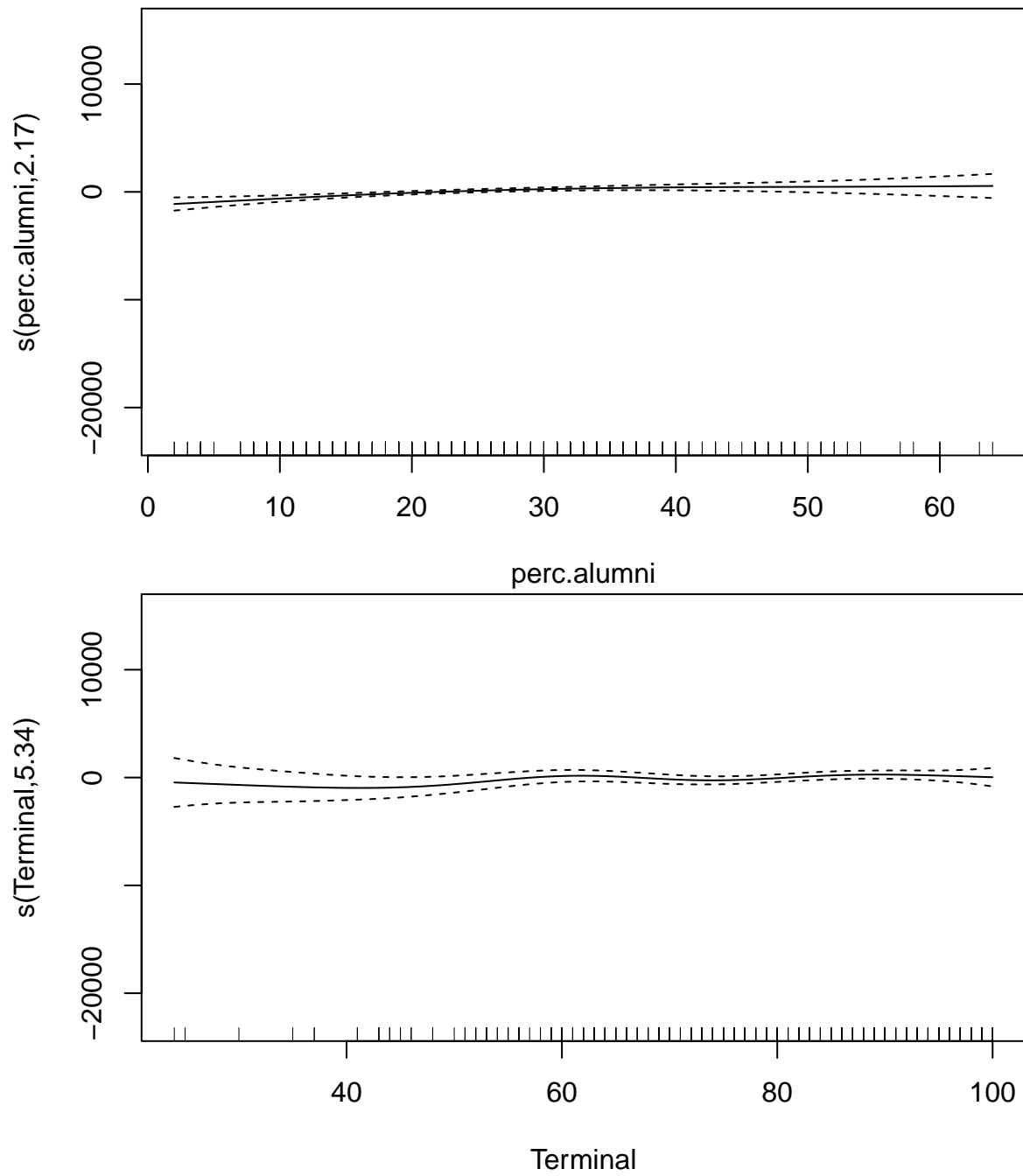
```
## select method
## 1 FALSE GCV.Cp
```

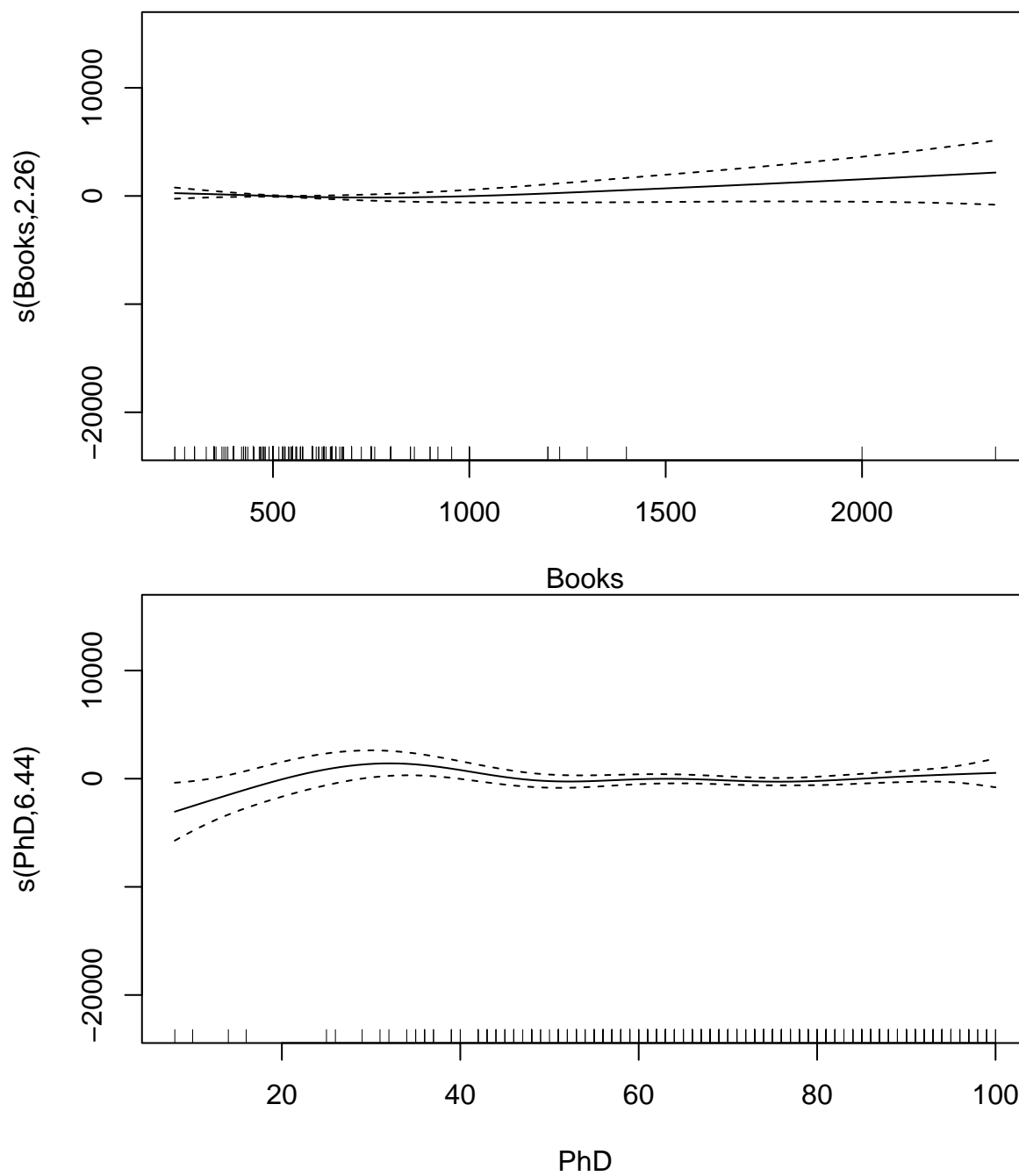
```
model.gam$finalModel
```

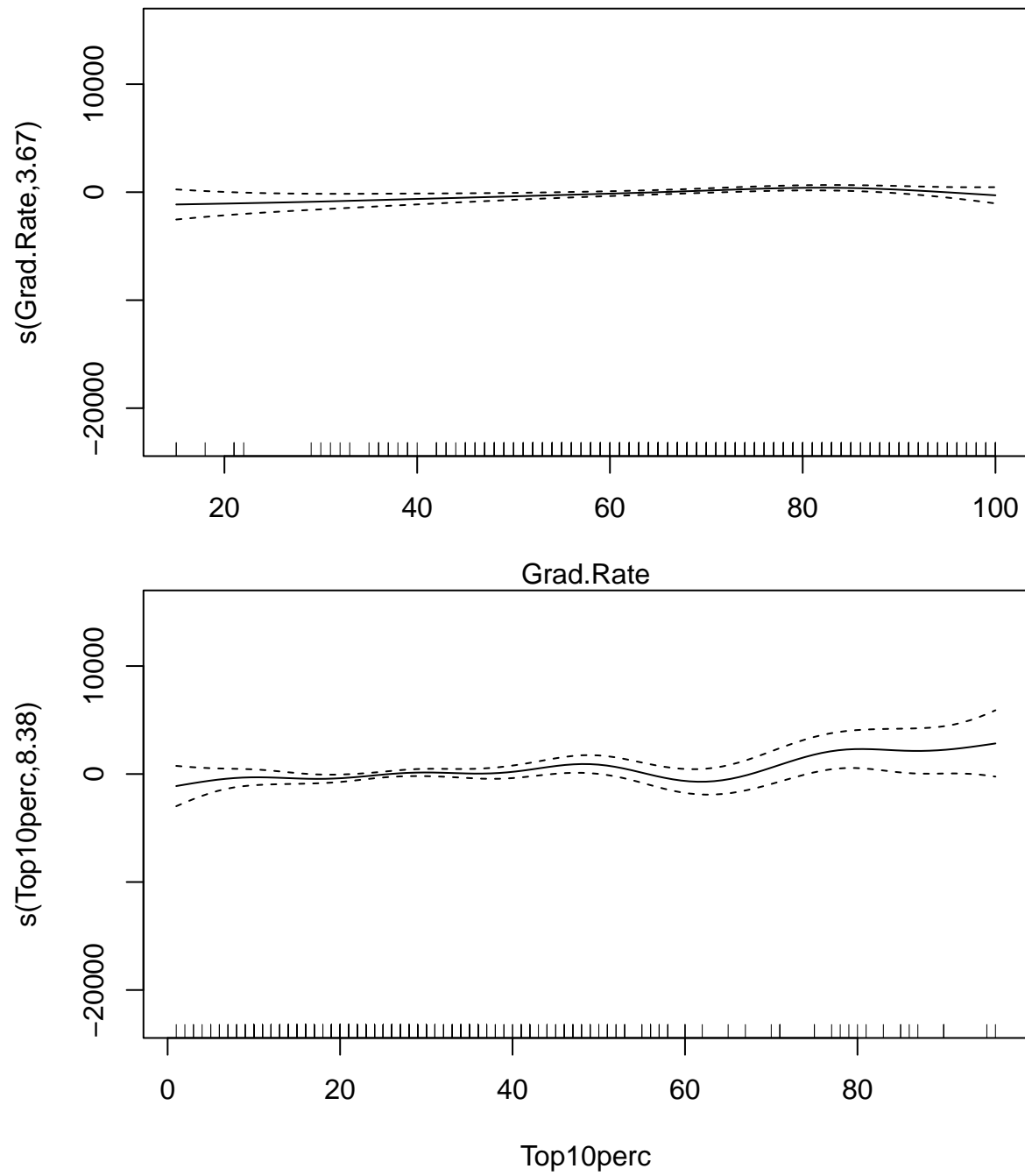
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(perc.alumni) + s(Terminal) + s(Books) + s(PhD) +
##      s(Grad.Rate) + s(Top10perc) + s(Top25perc) + s(S.F.Ratio) +
##      s(Personal) + s(P.Undergrad) + s(Enroll) + s(Room.Board) +
##      s(Accept) + s(Apps) + s(F.Undergrad) + s(Expend)
##
## Estimated degrees of freedom:
## 2.17 5.34 2.26 6.44 3.67 8.38 1.00
## 2.72 2.42 1.00 1.00 2.91 1.75 1.00
## 6.06 7.62 total = 56.75
##
## GCV score: 2839524
```

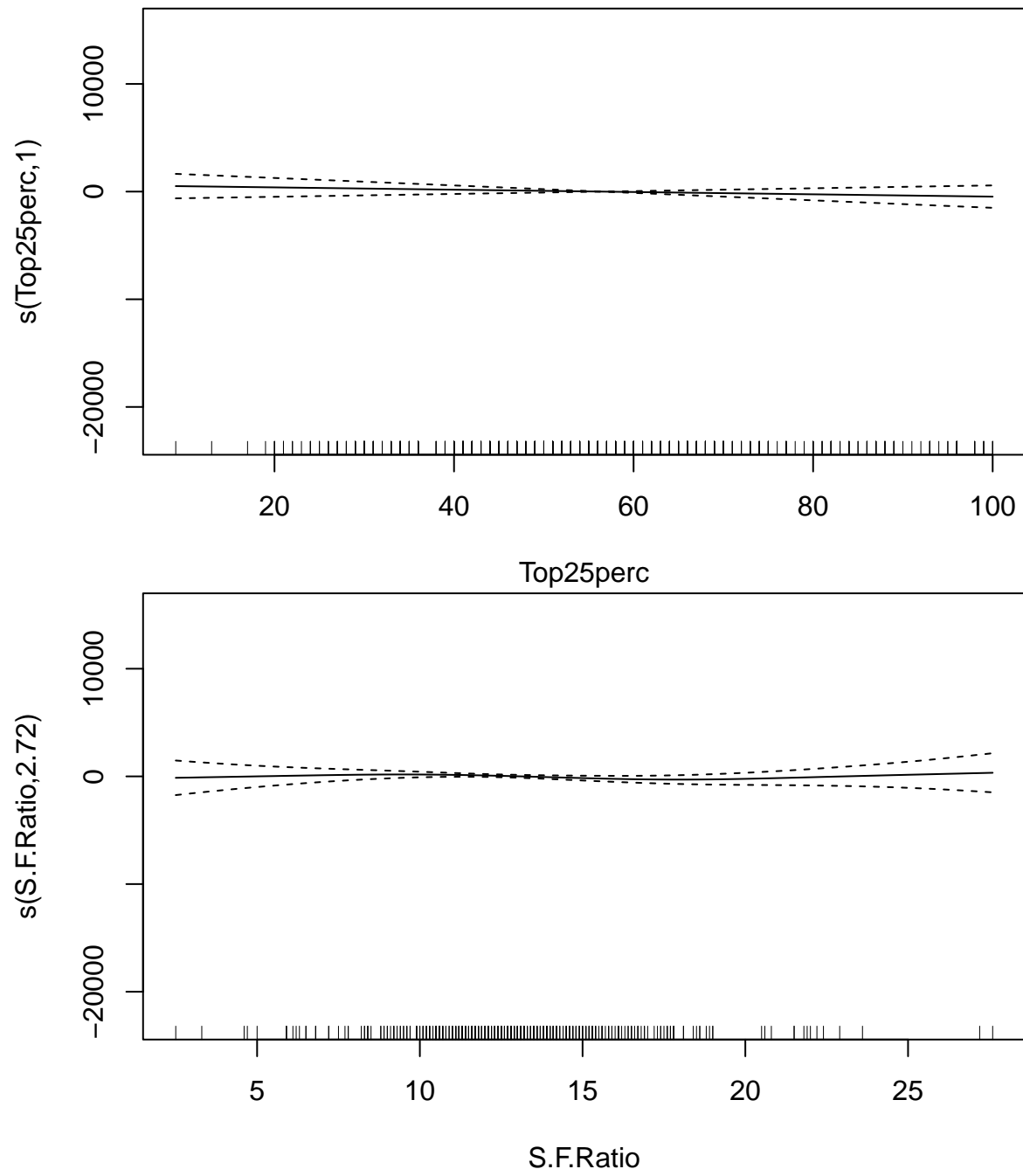
```
# degree of freedom=1 means linear
```

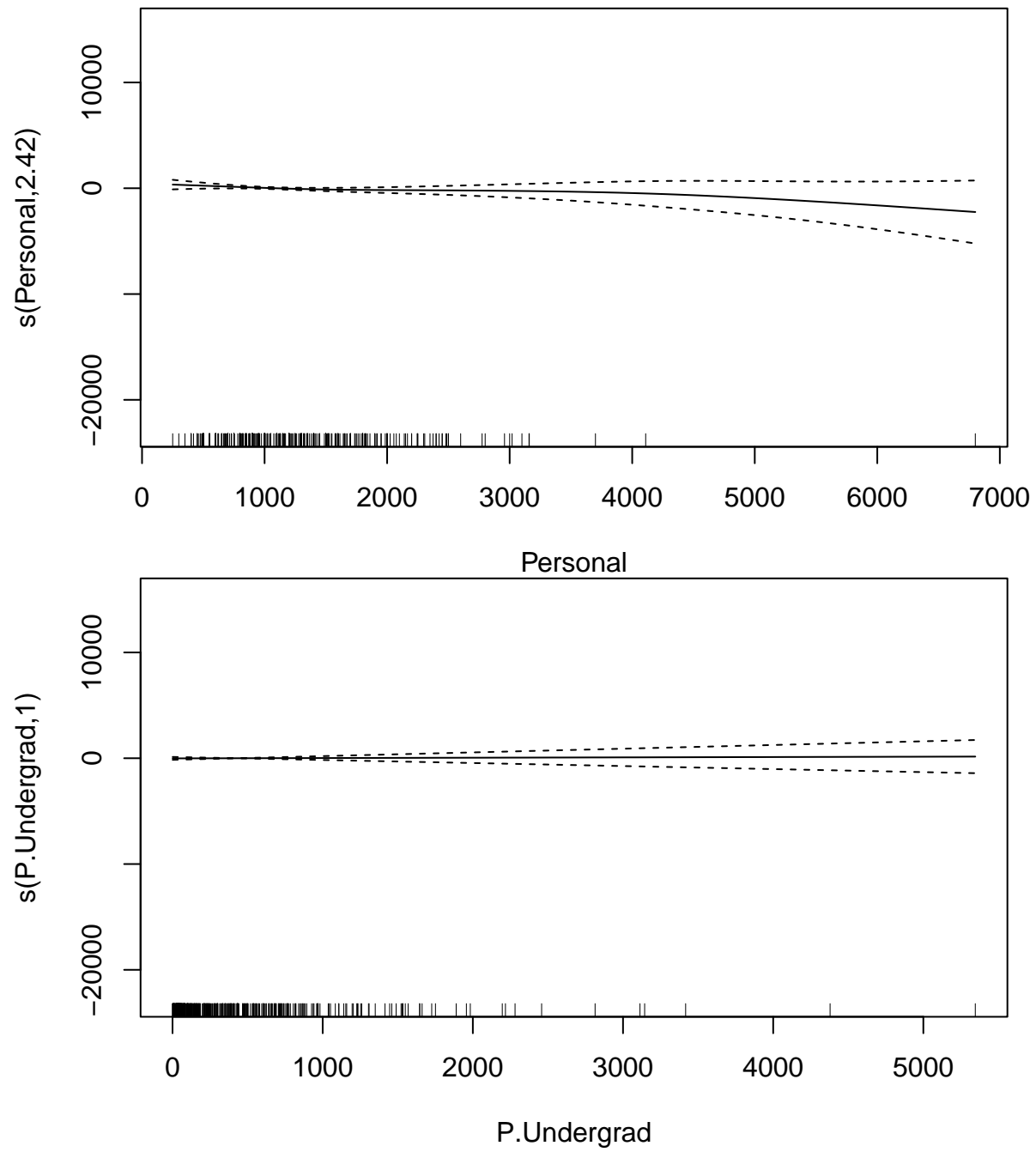
```
# Plotting
plot(model.gam$finalModel)
```

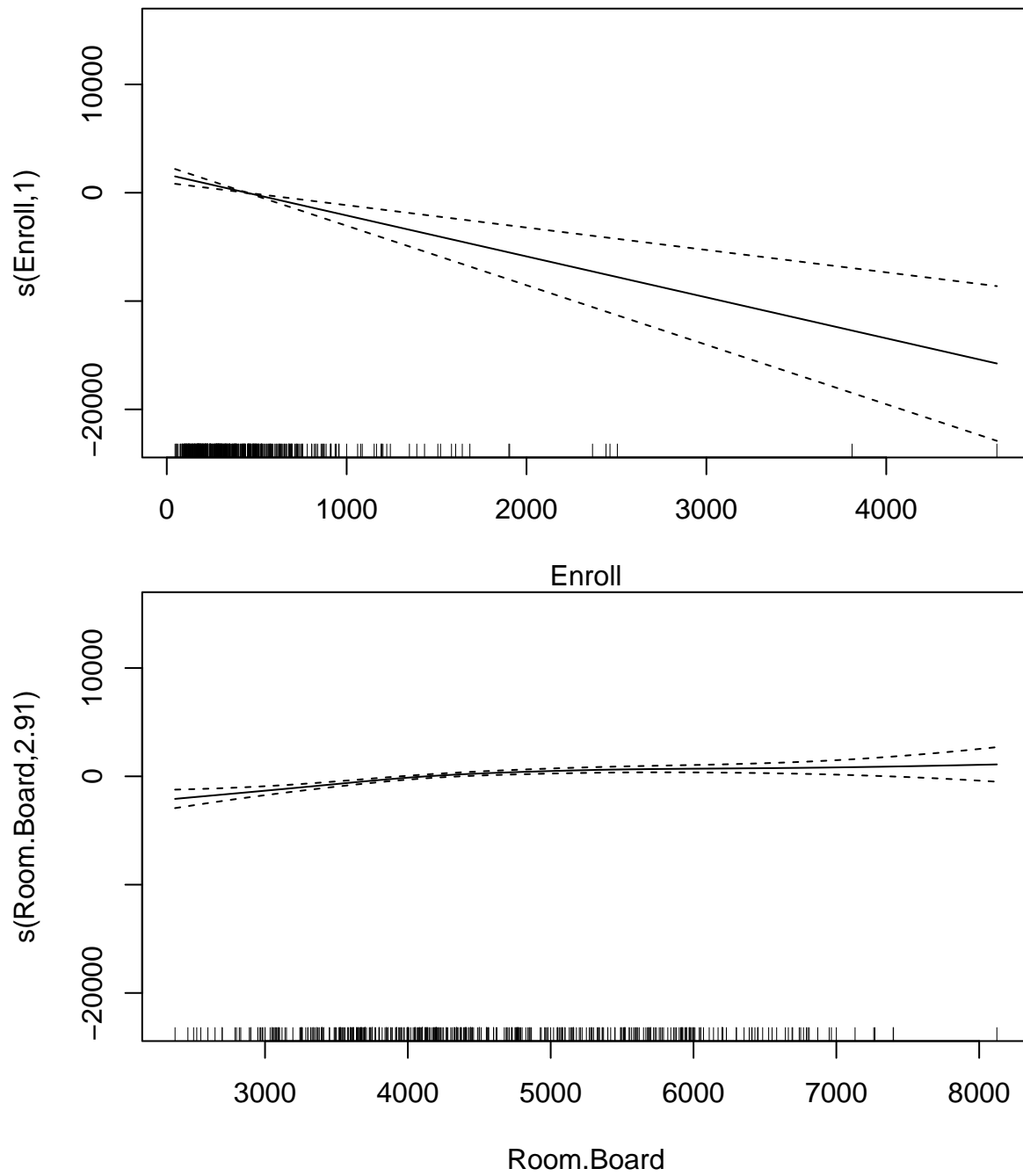


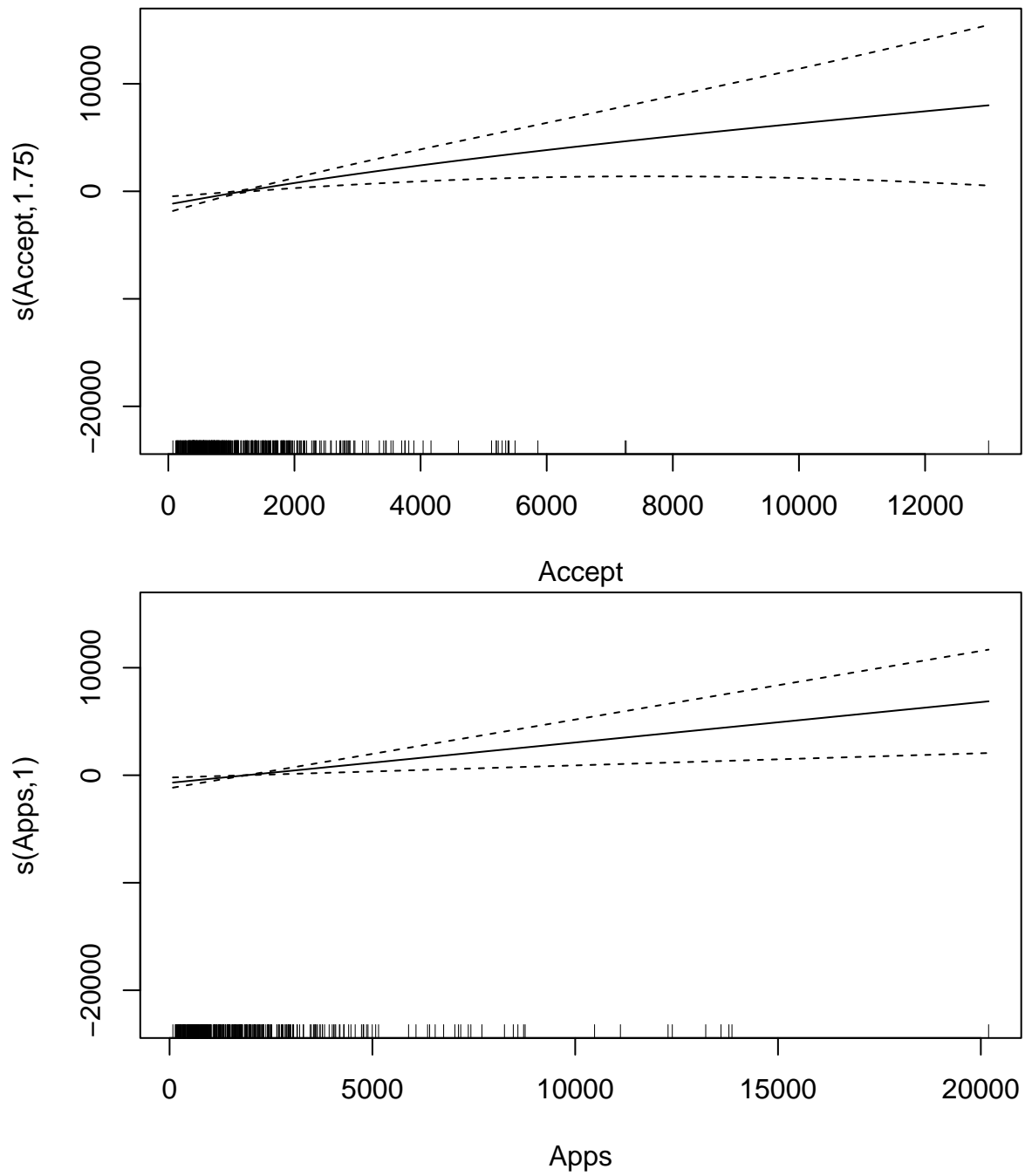


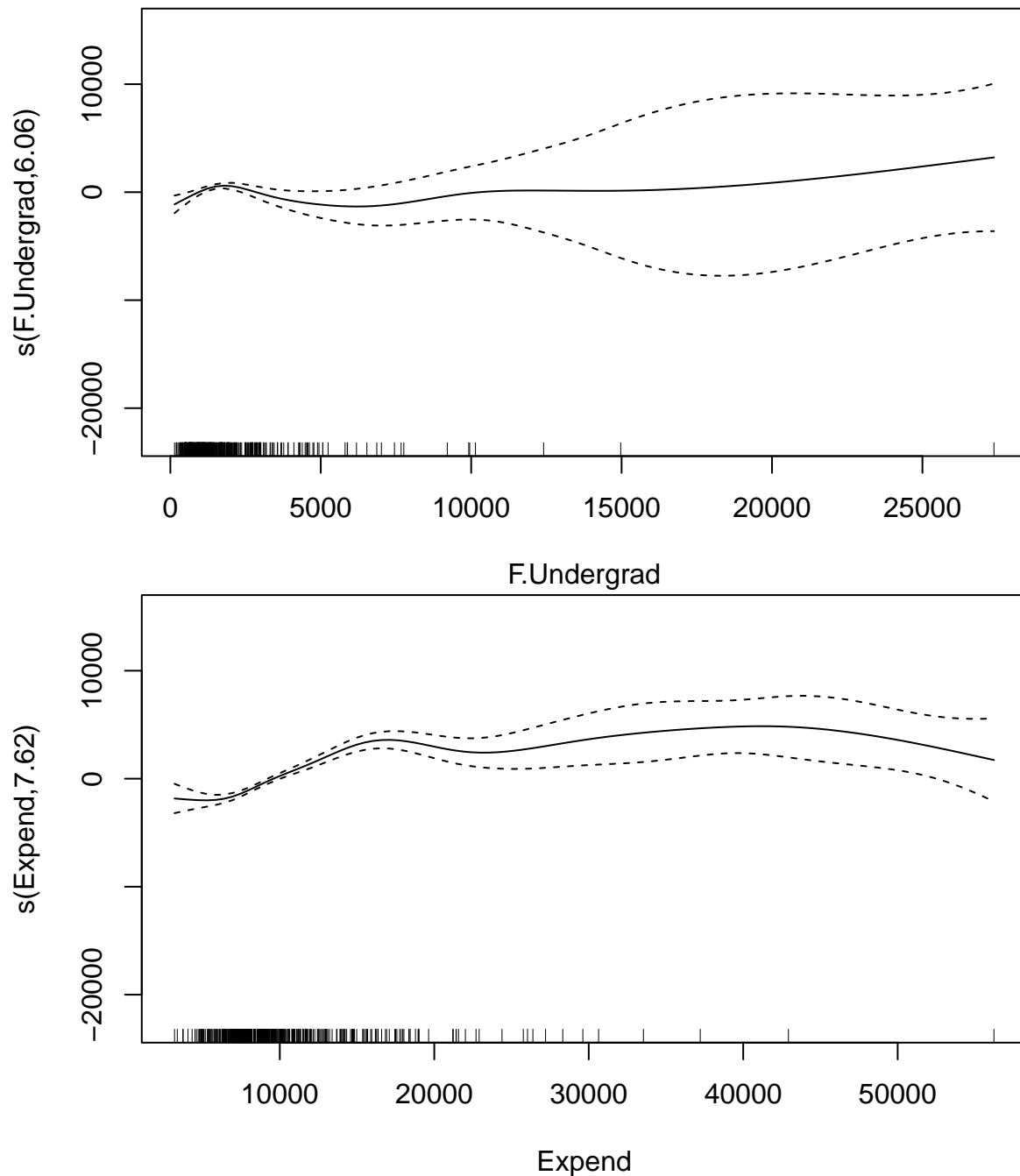












This GAM model includes all the predictors. The nonlinear terms included in the model are `perc.alumni`, `Terminal`, `Books`, `PhD`, `Grad.Rate`, `Top10perc`, `S.F.Ratio`, `Personal`, `P.Undergrad`, `S.F.Ratio`, `Personal`, `Room.Board`, `Accept`, `F.Undergrad`, `Expend`.

Discussion

For `perc.alumni`, `Terminal`, `Books`, `PhD`, `Grad.Rate`, `Top10perc`, `Top25perc`, `S.F.Ratio`, `Personal`, `P.Undergrad`, and `Room.Board`, the smoother appears relatively flat, suggesting that the relationship between these variables and the response variable is weak. For `Accept`, `Apps`, `F.Undergrad`, `Expend`, the smoothing spline increases as the number of applications increases, suggesting that the response variable tends to increase as these variables increases. For `Enroll`, the smoothing spline decreases as the number of applications increases, suggesting that the response variable tends to decrease as these variables increases.

```
# compute and report the test error
predictions <- predict(model.gam, x_test)
test_error <- mean((predictions - y_test)^2) # Mean Squared Error (MSE)
test_error # Reporting the test error
```

```
## [1] 2858273
```

The test error is 2.858273×10^6

(d)

In this dataset, would you favor a MARS model over a linear model for predicting out-of- state tuition? If so, why? More broadly, in general applications, do you consider a MARS model to be superior to a linear model? Please share your reasoning.

fit a linear model

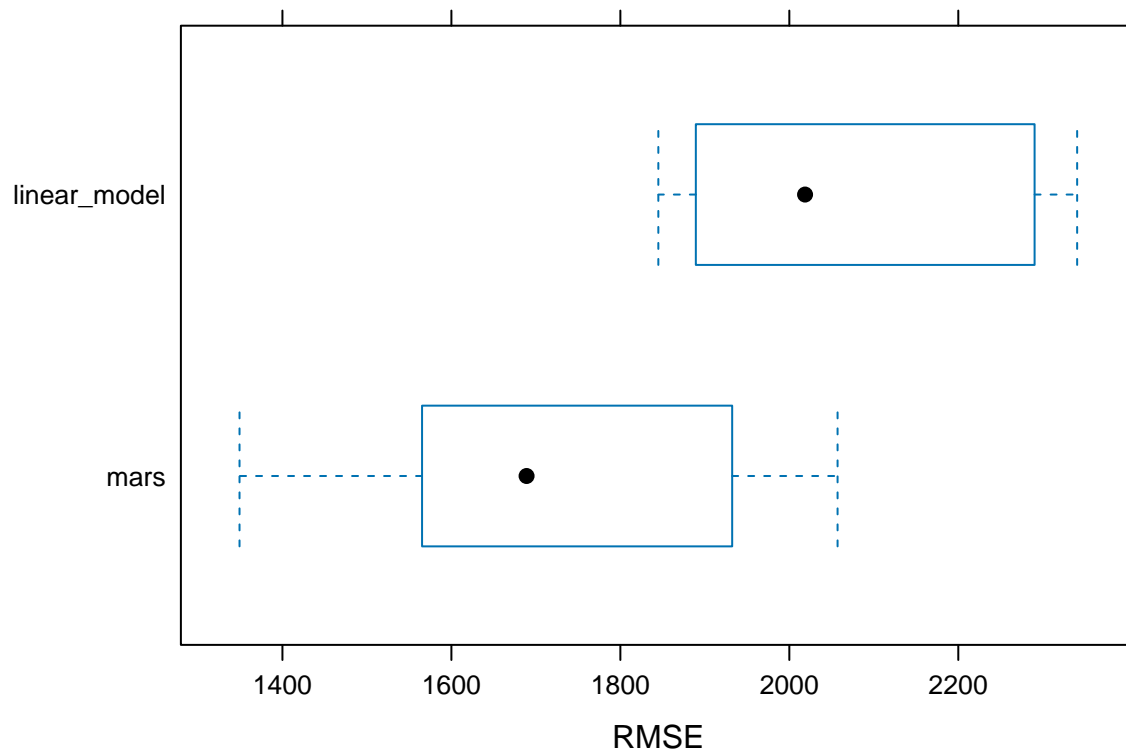
```
model.lm<-train(x_train, y_train,
               method = "lm",
               trControl = ctrl)
```

compare the RMSE

```
resamp <- resamples(list(linear_model = model.lm,
                        mars=model.mars))
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: linear_model, mars
## Number of resamples: 10
##
## MAE
##           Min.   1st Qu.   Median     Mean   3rd Qu.     Max. NA's
## linear_model 1458.544 1504.728 1597.661 1626.955 1700.601 1988.082    0
## mars         1111.145 1193.358 1346.749 1323.503 1385.608 1695.619    0
##
## RMSE
##           Min.   1st Qu.   Median     Mean   3rd Qu.     Max. NA's
## linear_model 1844.963 1902.257 2018.609 2064.771 2249.922 2340.698    0
## mars         1349.203 1569.036 1688.987 1723.255 1903.459 2057.105    0
##
## Rsquared
##           Min.   1st Qu.   Median     Mean   3rd Qu.     Max. NA's
## linear_model 0.5887687 0.6764265 0.7389193 0.7134490 0.7507739 0.8129015    0
## mars         0.7017853 0.7321717 0.7994049 0.7834563 0.8200695 0.8673229    0
```

```
bwplot(resamp, metric = "RMSE")
```



The MARS model is preferred since it has a lower mean value of RMSE compared to the linear model.

Whether a MARS model is superior to a linear model depends on the specifics of the data (the underlying true model).