p8106_hw2

Hao Zheng(hz2770)

2022/3/5

```
# Data Cleaning
dat =
  read.csv("./data/college.csv")[-1] %>%
  janitor::clean_names() %>%
  na.omit()

# Data Partition
indexTrain <- createDataPartition(y = dat$outstate, p = 0.8, list = FALSE)
trainData <- dat[indexTrain,]
testData <- dat[-indexTrain,]
head(trainData)
```

```
##   apps accept enroll top10perc top25perc f_undergrad p_undergrad outstate
## 1 1660   1232    721        23        52        2885         537     7440
## 2 2186   1924    512        16        29        2683        1227    12280
## 3 1428   1097    336        22        50        1036          99    11250
## 4  417    349    137        60        89         510          63    12960
## 5  193    146     55        16        44         249         869     7560
## 7  353    340    103        17        45         416         230    13290
##   room_board books personal ph_d terminal s_f_ratio perc_alumni expend
## 1       3300   450     2200   70       78      18.1          12   7041
## 2       6450   750     1500   29       30      12.2          16  10527
## 3       3750   400     1165   53       66      12.9          30   8735
## 4       5450   450      875   92       97       7.7          37  19016
## 5       4120   800     1500   76       72      11.9           2  10922
## 7       5720   500     1500   90       93      11.5          26   8861
##   grad_rate
## 1        60
## 2        56
## 3        54
## 4        59
## 5        15
## 7        63
```

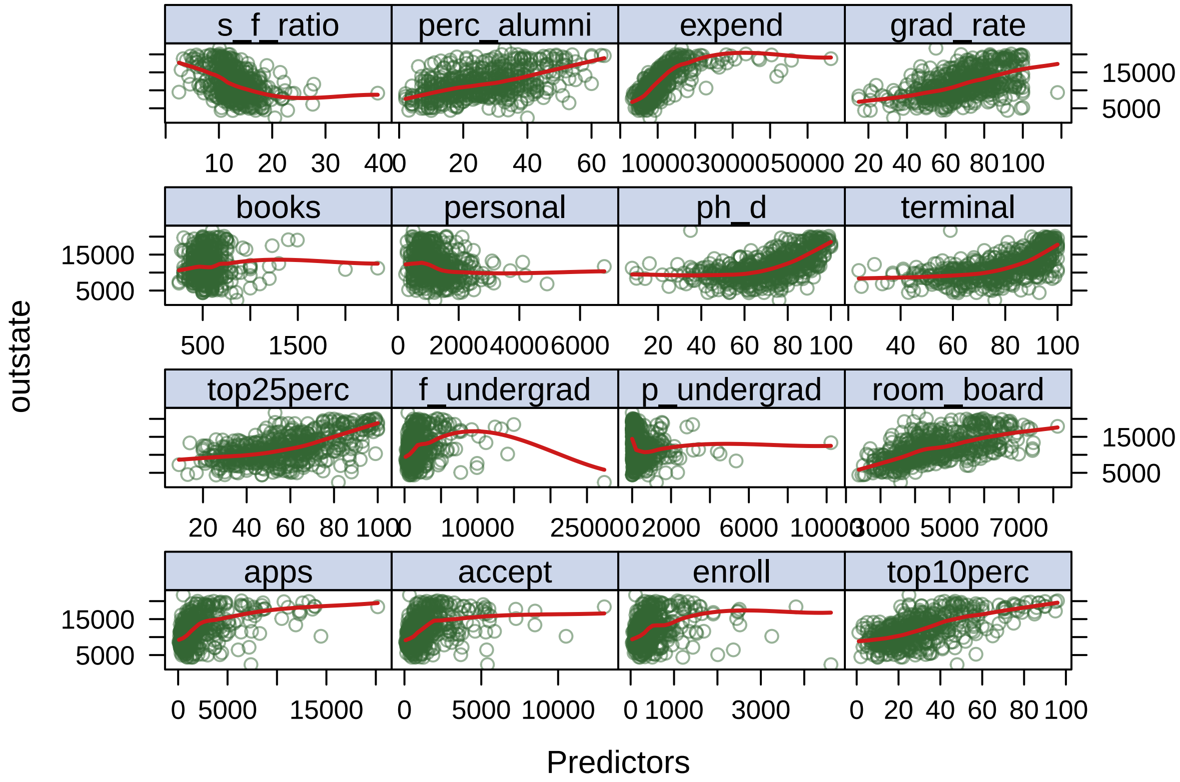
Exploratory Data Analysis

```
theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$psh <- 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)

x <- dat %>% select(-outstate)
y <- dat$outstate

# scatter plot
featurePlot(x,
            y,
            plot = "scatter",
            span = .5,
            labels = c("Predictors", "outstate"),
            type = c("p", "smooth"),
            layout = c(4,4))
```



From the scatter plot, we can see that most predictors are not linearly associated with the response variable. However, there may exist a linear relationship between the variable `perc_alumni`, `grad_rate`, `room_board` and the response `outstate` respectively.

Smoothing Spline Models

Now let's fit smoothing spline models using `terminal` as the only predictor of `outstate`.

```
terminal.grid <- seq(from = 40, to = 100, by = 10)
fit.ss <- smooth.spline(trainData$terminal, trainData$outstate)
fit.ss$df
```

```
## [1] 4.26278
```

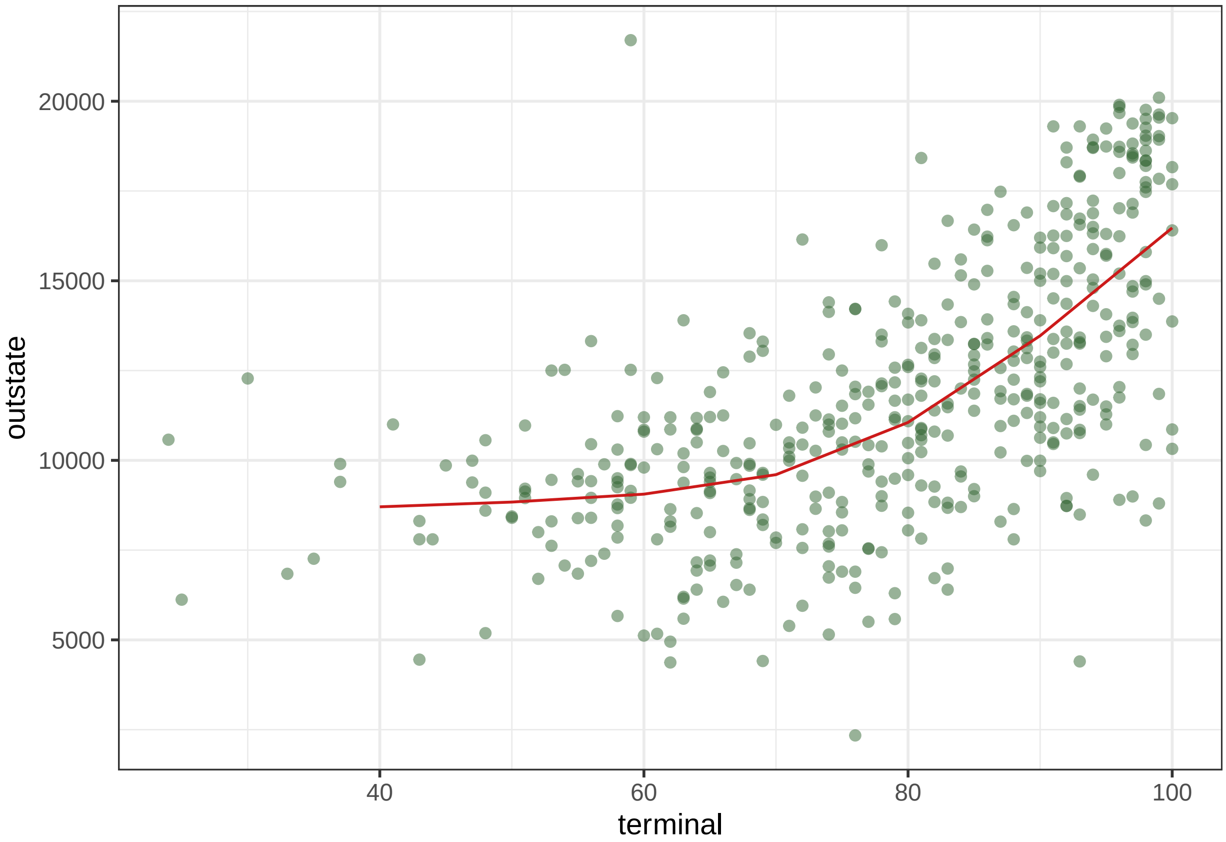
```
fit.ss$lambda
```

```
## [1] 0.0412237
```

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

# plot the fit
p <- ggplot(data = trainData, aes(x = terminal, y = outstate)) +
  geom_point(color = rgb(.2, .4, .2, .5))

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



The smoothing spline model fitted for a range of degrees of freedom is 4.2627796. Then obtain the degrees of freedom using generalized cross-validation and plot the new fits.

```
fit.ss.cv <- smooth.spline(trainData$terminal, trainData$outstate, cv = TRUE)
```

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

```
fit.ss.cv$df
```

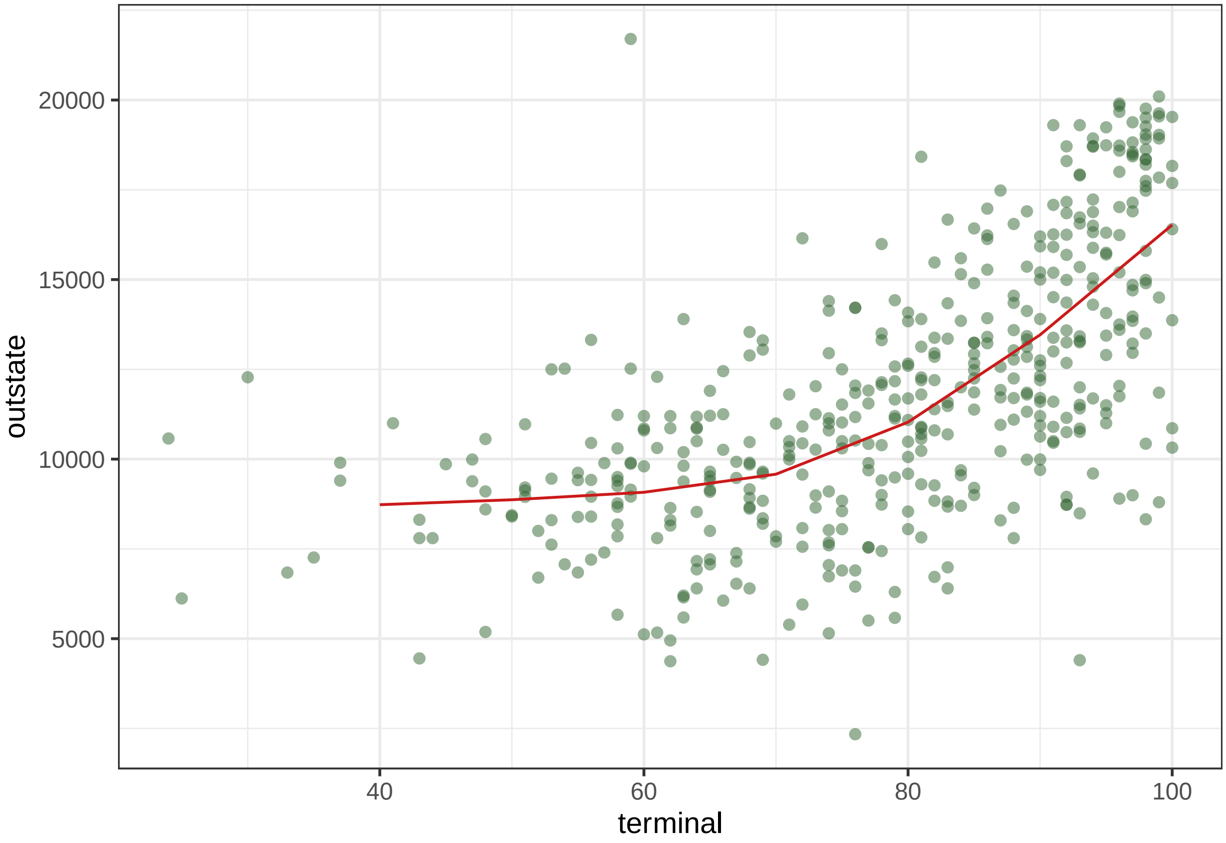
```
## [1] 4.492168
```

```
fit.ss.cv$lambda
```

```
## [1] 0.03175134
```

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

p +
  geom_line(aes(x = terminal.grid, y = pred), data = pred.ss.df.cv, color = rgb(.8, .1, .1, 1)) + theme_
```



Using cross-validation, we obtain the degrees of freedom 4.4921683 with lambda = 0.0317513.

Generalized Additive Models (GAM)

Multivariate Adaptive Regression Spline (MARS)

Model Selection