# p8106_hw2_qz2266

Qing Zhou

2023-03-10

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'tidyr' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

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

```
## Warning: package 'mgcv' was built under R version 4.2.2
```

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

```
## Warning: package 'nlme' was built under R version 4.2.2
```

```
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:dplyr':
##
##     collapse
##
## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.
```

```
library(earth)
```

```
## Warning: package 'earth' was built under R version 4.2.2
```

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

```
## Warning: package 'plotmo' was built under R version 4.2.2
```

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

```
## Warning: package 'TeachingDemos' was built under R version 4.2.2
```

```
library(pdp)
```

```
## Warning: package 'pdp' was built under R version 4.2.2
```

```
##
## Attaching package: 'pdp'
##
## The following object is masked from 'package:purrr':
##
##     partial
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

## Data preparation

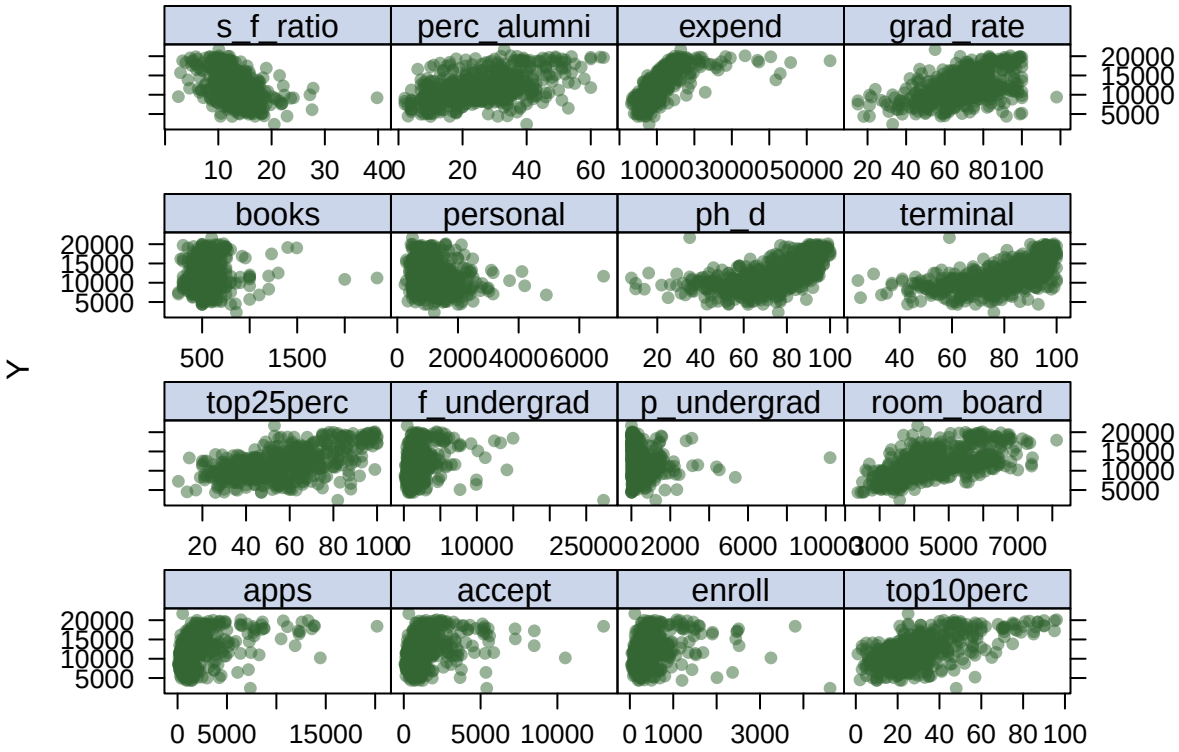
In this exercise, we build nonlinear models using the "College" data. The dataset contains statistics for 565 US Colleges from a previous issue of US News and World Report. The response variable is the out-of-state tuition (Outstate).

We exclude the statistics of Columbia University (i.e., the 125th observation) to train the models.

```
college_df = read_csv('./data/College.csv') %>%
  janitor::clean_names()

  #skimr::skim(college_df)

college_train = college_df %>%
  filter(college != "Columbia University") %>%
  select(-college)
```

## Exploratory data analysis

```
# matrix of predictors
x = model.matrix(outstate ~ ., college_train)[,-1]
# vector of response
y = college_train$outstate

# the relationship of response vs. predictors
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)

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


s_f_ratio
perc_alumni
expend
grad_rate
books
personal
ph_d
terminal
top25perc
f_undergrad
p_undergrad
room_board
apps
accept
enroll
top10perc
Y