

# P8106\_\_stl2137\_\_HW2

The response variable is the out-of-state tuition (Outstate).

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

```
## Parsed with column specification:  
## cols(  
##   College = col_character(),  
##   Apps = col_double(),  
##   Accept = col_double(),  
##   Enroll = col_double(),  
##   Top10perc = col_double(),  
##   Top25perc = col_double(),  
##   F.Undergrad = col_double(),  
##   P.Undergrad = col_double(),  
##   Outstate = col_double(),  
##   Room.Board = col_double(),  
##   Books = col_double(),  
##   Personal = col_double(),  
##   PhD = col_double(),  
##   Terminal = col_double(),  
##   S.F.Ratio = col_double(),  
##   perc.alumni = col_double(),  
##   Expend = col_double(),  
##   Grad.Rate = col_double()  
## )
```

```
school_no_columbia_dat <- school_dat[-125,]
```

## Part A

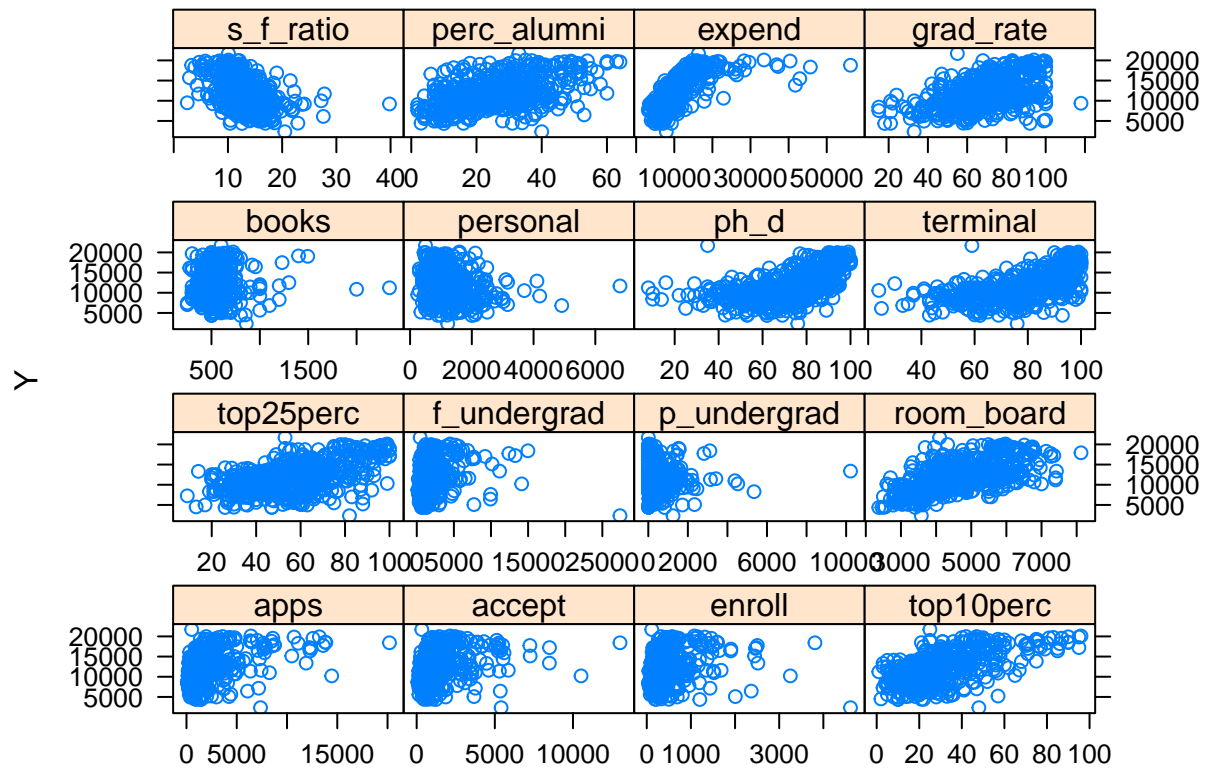
```
predictors_df <- school_no_columbia_dat %>%  
  select(outstate, everything()) %>%  
  group_by(college) %>%  
  pivot_longer(  
    apps:grad_rate,  
    names_to = "predictor",  
    values_to = "values"  
  )  
  
predictors_df %>%  
  ggplot(  
    aes(x = values, y = outstate, color = predictor)  
  ) +  
  geom_point(alpha = 0.25) +  
  facet_wrap(. ~ predictor, ncol = 4)
```



```
# matrix of predictors
school_no_columbia_dat <- school_no_columbia_dat[,-1]

x <- model.matrix(outstate~.,school_no_columbia_dat)[,-1]
y <- school_no_columbia_dat$outstate

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



## Part B

- (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.

```
smooth_spline_fit <- smooth.spline(school_no_columbia_dat$terminal, school_no_columbia_dat$outstate)
smooth_spline_fit$df

## [1] 4.468629

terminal_lims <- range(school_no_columbia_dat$terminal)
terminal_grid <- seq(from = terminal_lims[1], to = terminal_lims[2])

pred_smooth <- predict(smooth_spline_fit,
                      x = terminal_grid)

pred_sspline_df <- data.frame(pred = pred_smooth$y,
                             terminal = terminal_grid)

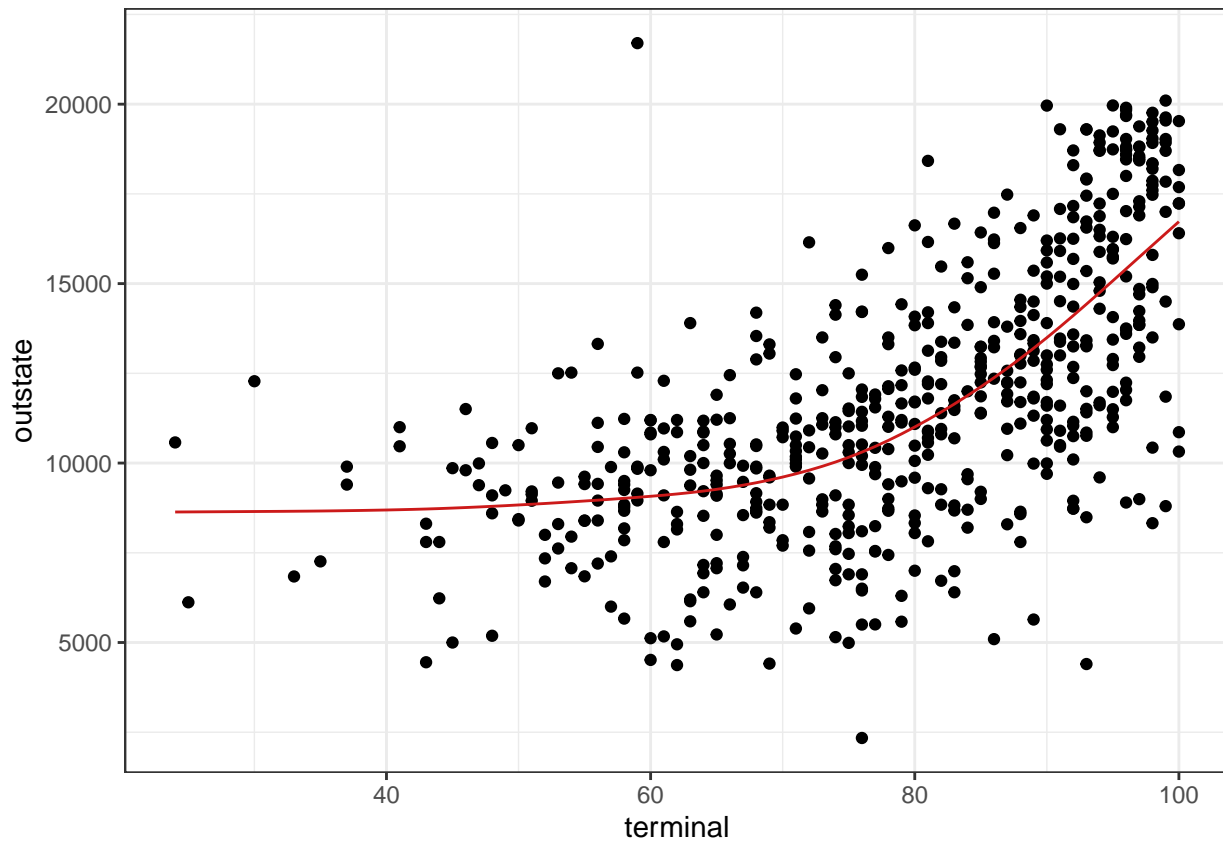
p <- ggplot(data = school_no_columbia_dat,
            aes(
              x = terminal,
              y = outstate
            )) + geom_point() + theme_bw()

p + geom_line(
  aes(
    x = terminal,
```

```

y = pred),
data = pred_sspline_df,
color = rgb(.8, .1, .1, 1)) + theme_bw()

```



From the smoothing spline model, we are able to ascertain that the degree of freedoms is 4.4686294.

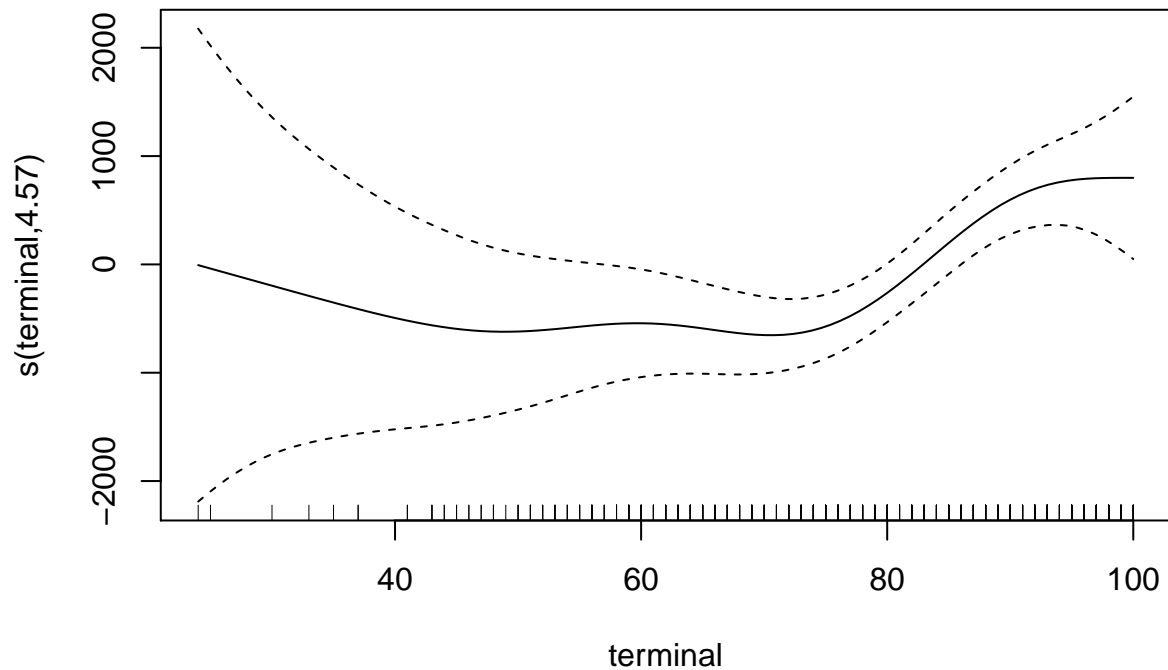
## Part C

(c) Plot the results and explain your findings.

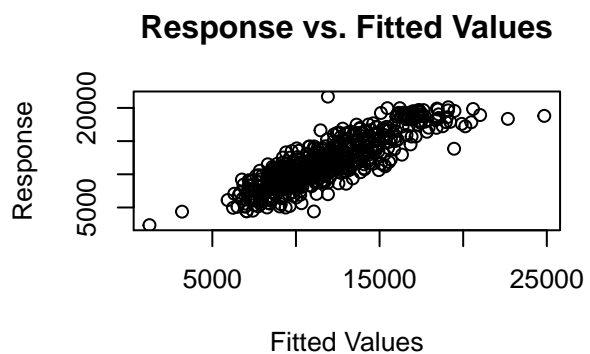
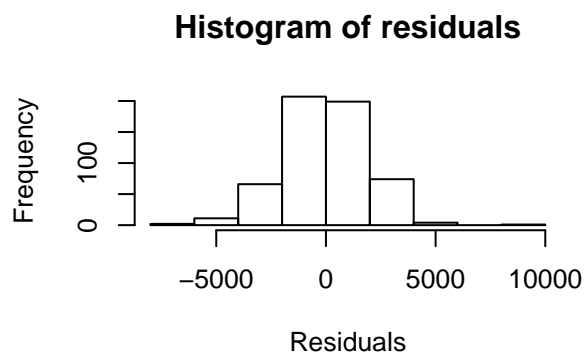
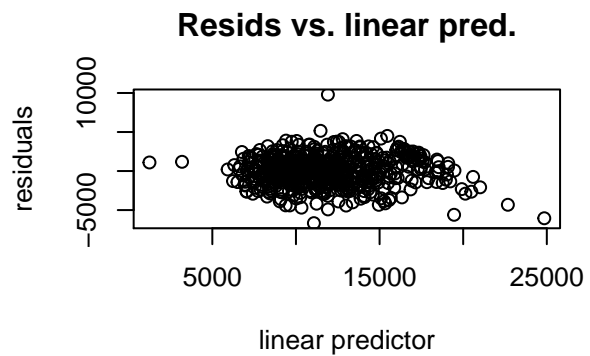
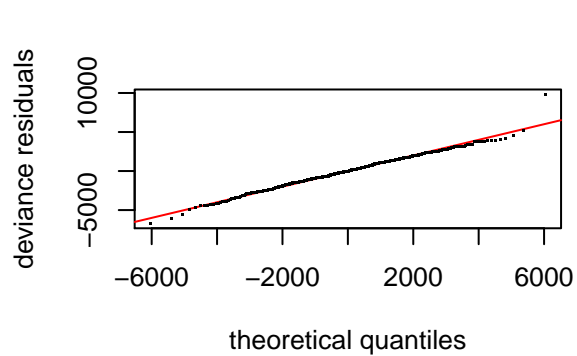
```

gam_school_1 <- gam(outstate ~ apps + accept + enroll + top10perc + top25perc + f_undergrad + p_undergrad)
plot(gam_school_1)

```



```
gam.check(gam_school_1)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 5 iterations.
## The RMS GCV score gradient at convergence was 45.19481 .
## The Hessian was positive definite.
## Model rank = 25 / 25
```

```
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##           k'   edf k-index p-value
## s(terminal) 9.00 4.57    0.95    0.14
```

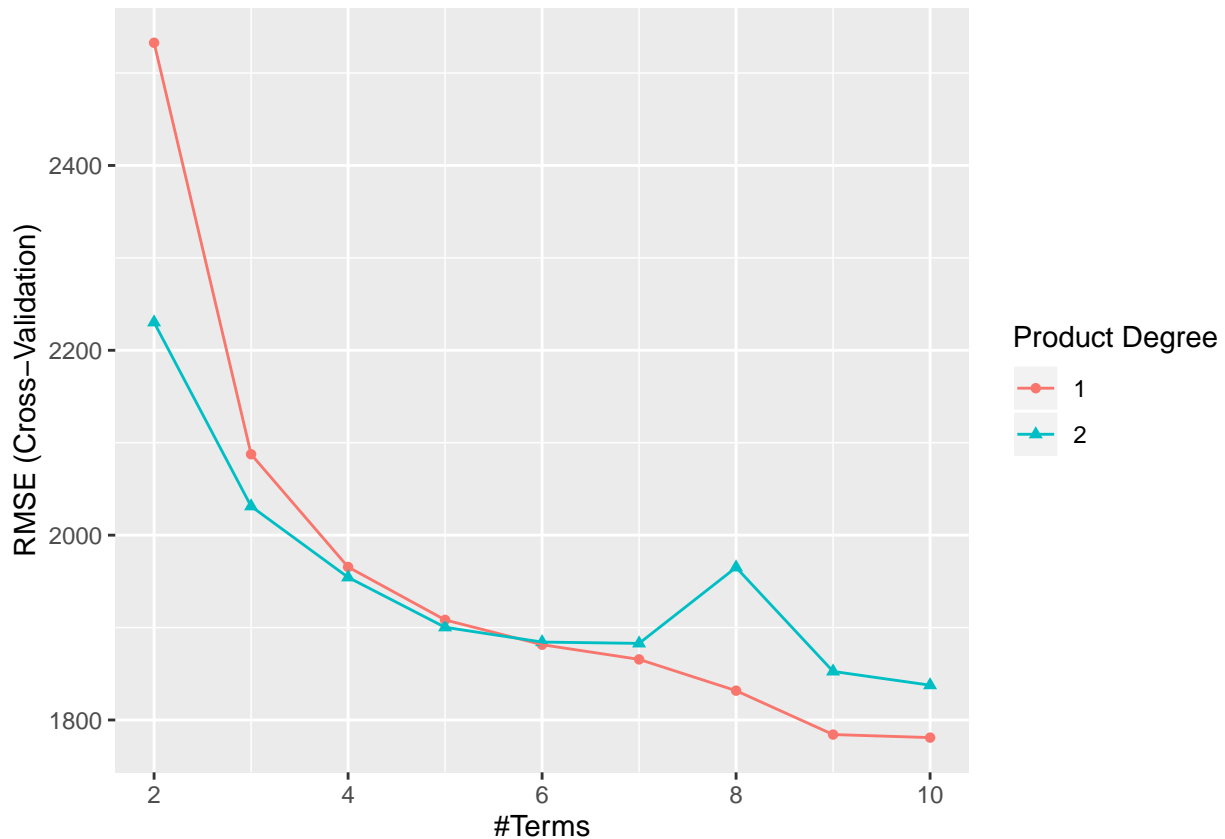
## Part D

```
mars_grid <- expand.grid(degree = 1:2,
                        nprune = 2:10)

control1 <- trainControl(method = "cv", number = 10)

set.seed(1)
mars_fit <- train(x, y,
                 method = "earth",
                 tuneGrid = mars_grid,
                 trControl = control1)

ggplot(mars_fit)
```



```
mars_fit$bestTune
```

```
##   nprune degree
## 9      10      1
```

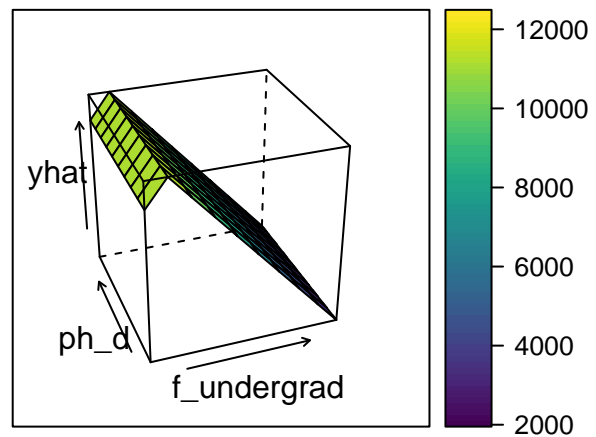
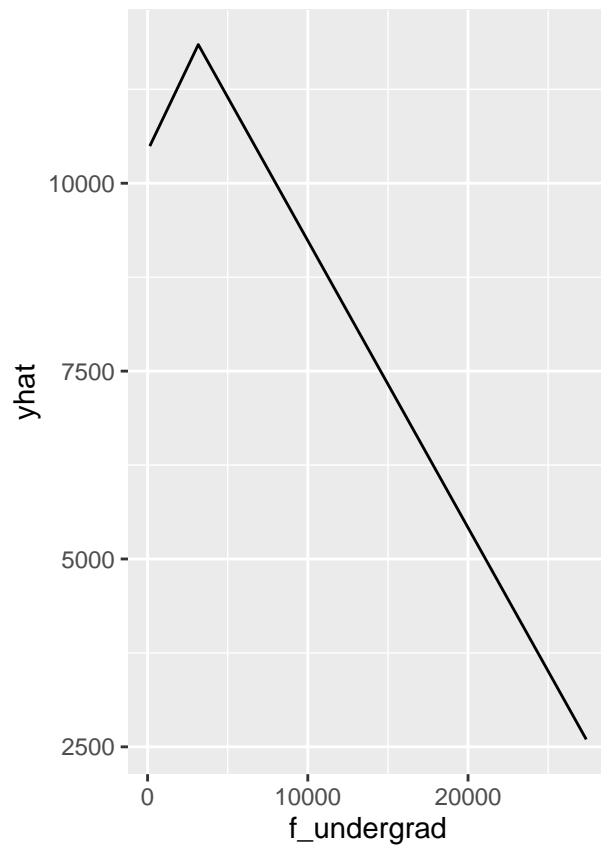
```
coef(mars_fit$finalModel)
```

```
##      (Intercept)      h(expend-15365) h(4450-room_board)
##      10856.8275542      -0.7836173      -1.4272043
## h(f_undergrad-1355) h(1355-f_undergrad) h(22-perc_alumni)
##      -0.3818847      -1.6799143      -105.5570689
##      h(apps-3712)      h(913-enroll)      h(2193-accept)
##      0.4334737      4.5019587      -1.9769988
##      h(expend-6881)
##      0.7774546
```

```
partial_school_1 <- partial(mars_fit, pred.var = c("f_undergrad"), grid.resolution = 10) %>% autoplot()
```

```
partial_school_2 <- partial(mars_fit, pred.var = c("f_undergrad", "ph_d"), grid.resolution = 10) %>%
  plotPartial(levelplot = FALSE, zlab = "yhat", drape = TRUE,
    screen = list(z = 20, x = -60))
```

```
grid.arrange(partial_school_1, partial_school_2, ncol = 2)
```



## Part E

```
### Grabbing Columbia observation
columbia_dat <- school_dat[125,]

columbia_gam <- predict(gam_school_1, newdata = columbia_dat)
```

```
columbia_mars <- as.numeric(predict(mars_fit, newdata = columbia_dat))
```

Based off the GAM model, we predict that the out-of-state tuition at Columbia University is  $1.9406713 \times 10^4$ . Based off the MARS model, we predict that the out-of-state tuition at Columbia University is  $1.7469904 \times 10^4$ . Between the two models, the GAM model predicts the out-of-state tuition for Columbia to be higher by 1936.8090809 compared to MARS model.