Homework 2

Jeffrey LIANG

2/20/2021

```r
set.seed(123123)
```

Q1

Table 1: Data summary

Name	clg_data
Number of rows	565
Number of columns	18
Column type frequency:	
factor	1
numeric	17
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
college	0	1	FALSE	565	Abi: 1, Ade: 1, Adr: 1, Agn: 1

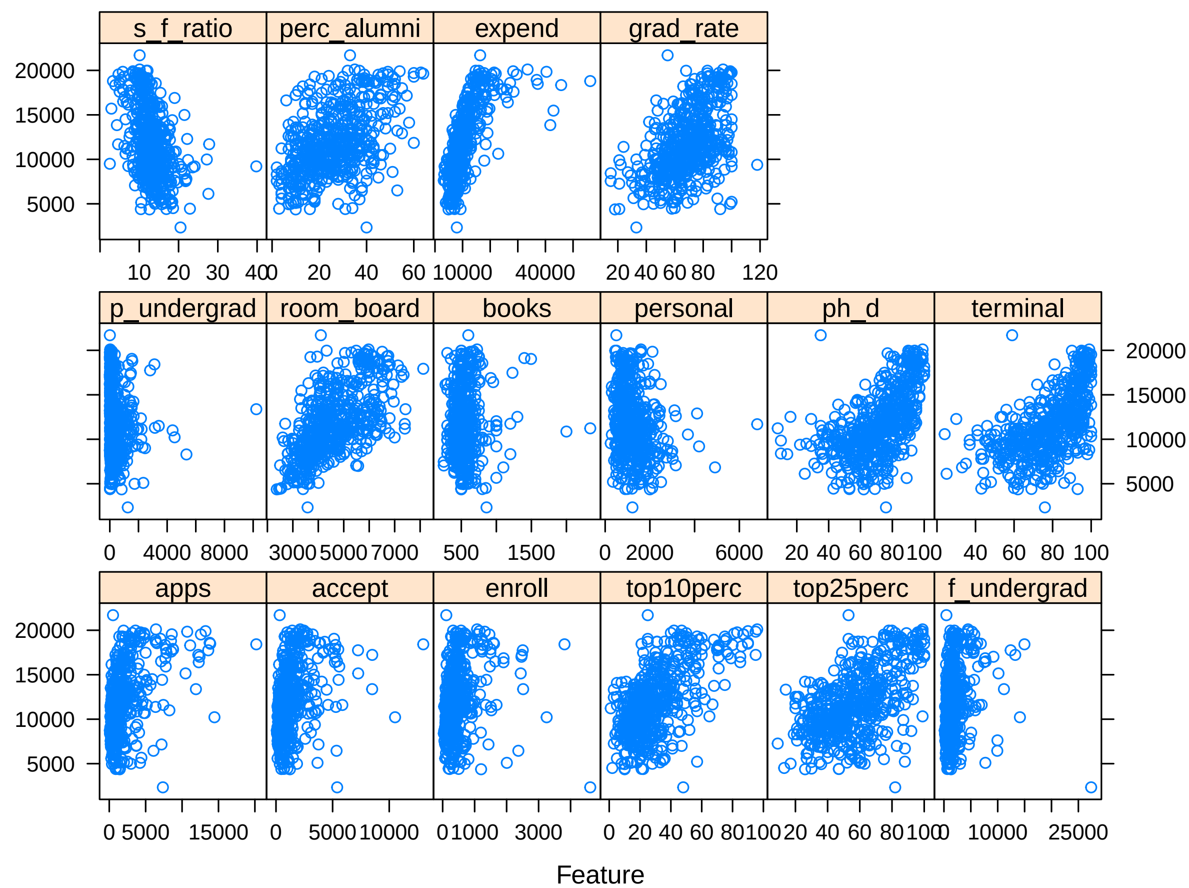
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
apps	0	1	1977.9	2443.34	81.0	619.0	1133.0	2186.0	20192.0
accept	0	1	1305.7	1369.55	72.0	501.0	859.0	1580.0	13007.0
enroll	0	1	456.9	457.53	35.0	206.0	328.0	520.0	4615.0
top10perc	0	1	29.3	17.85	1.0	17.0	25.0	36.0	96.0
top25perc	0	1	57.0	19.59	9.0	42.0	55.0	70.0	100.0
f_undergrad	0	1	1872.2	2110.66	139.0	840.0	1274.0	2018.0	27378.0
p_undergrad	0	1	434.0	722.37	1.0	63.0	207.0	541.0	10221.0
outstate	1	1	11789.6	3699.59	2340.0	9100.0	11200.0	13962.5	21700.0
room_board	0	1	4586.1	1089.70	2370.0	3736.0	4400.0	5400.0	8124.0
books	0	1	547.5	174.93	250.0	450.0	500.0	600.0	2340.0
personal	0	1	1214.4	632.88	250.0	800.0	1100.0	1500.0	6800.0
ph_d	0	1	71.1	17.35	8.0	60.0	73.0	85.0	100.0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
terminal	0	1	78.5	15.45	24.0	68.0	81.0	92.0	100.0
s_f_ratio	0	1	12.9	3.52	2.5	11.1	12.7	14.5	39.8
perc_alumni	0	1	25.9	12.40	2.0	16.0	25.0	34.0	64.0
expend	0	1	10486.4	5682.58	3186.0	7477.0	8954.0	11625.0	56233.0
grad_rate	0	1	69.0	16.75	15.0	58.0	69.0	81.0	118.0

Missing data is the respone, omitting the data instead of treating with data preprocessing.

```
clg_data = clg_data %>% drop_na()

clg_train = clg_data

Y_train = clg_train$outstate

X_train = model.matrix(outstate ~., data = clg_train)[,-1]

ctrl = trainControl(method = "repeatedcv",number = 5, repeats = 5)

clg_data %>%
  select(-college,-outstate) %>%
  featurePlot(.,clg_data$outstate,plot = "scatter",row = 4)
```



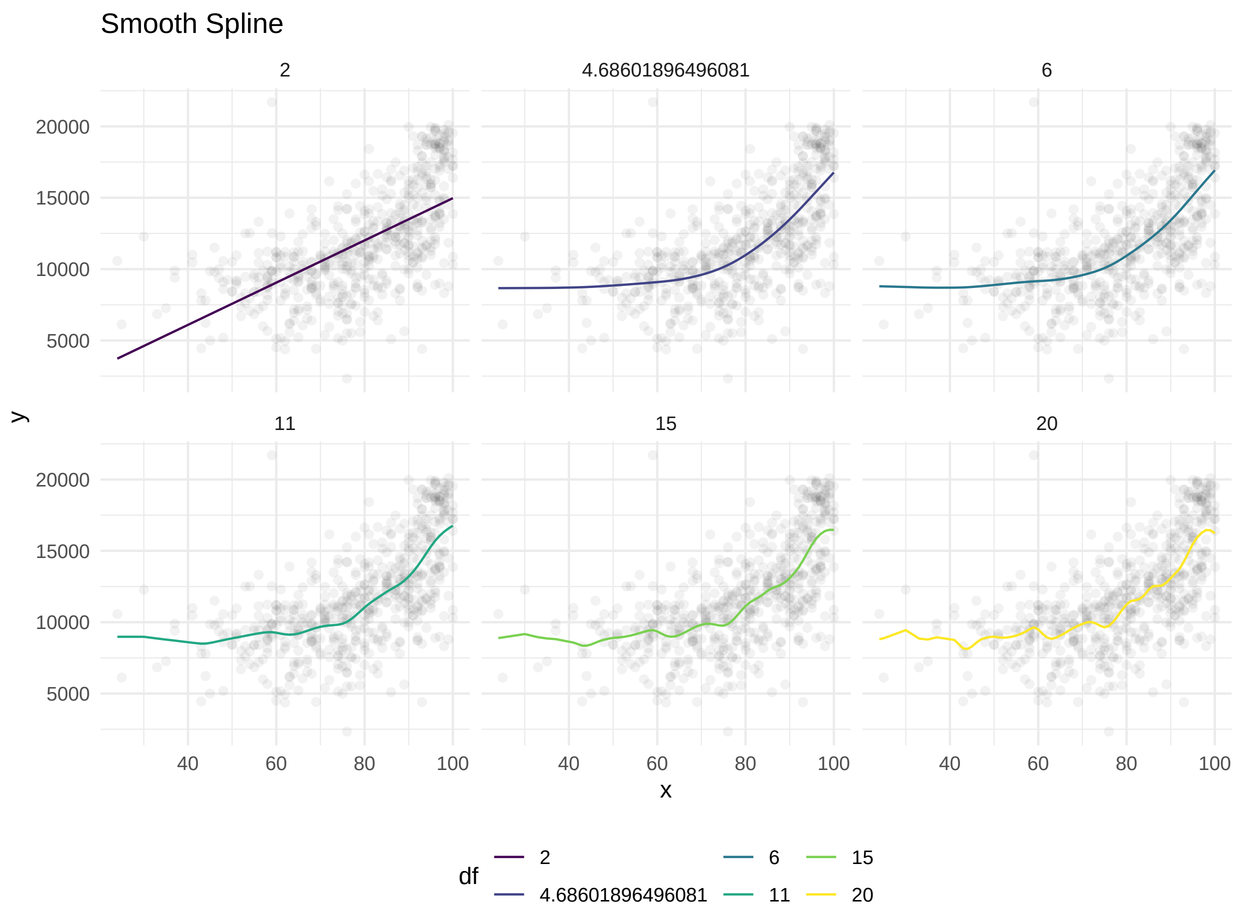
Q2

```
set.seed(123123)
clg_ss_cv = smooth.spline(clg_train$terminal, Y_train, cv = T)

clg_ss_cv_mse = mean((predict(clg_ss_cv,clg_train$terminal,se=F)$y-Y_train)^2)

clg_ss =
  tibble(
    x = list(clg_train$terminal),
    y = list(Y_train),
    df = list(seq(2, 20, length = 5)%/%1)
  ) %>%
  unnest(df) %>%
  mutate(model = pmap(list(x, y, df),
                      function(x, y, df, ...)
                        smooth.spline(
                          x = x, y = y, df = df
                        ))) %>%
  rbind(list(
    x = list(clg_train$terminal),
    y = list(Y_train),
    df = clg_ss_cv$df,
    model = list(clg_ss_cv)
  )) %>%
  mutate(
    prediction = map2(.x = x,
                      .y = model,
                      ~predict(object = .y,x = .x,se=F)$y),
    df = as.factor(df)
  ) %>%
  select(df, y, prediction, x) %>%
  unnest(c(prediction, y,x))

clg_ss %>%
  group_by(df) %>%
  summarise(mse =
            mean((y - prediction) ^ 2)) %>%
  knitr::kable(caption = "Smooth spline performance with different degree of freedom",digits = 3)
```

Table 4: Smooth spline performance with different degree of freedom

df	mse
2	8449920
4.68601896496081	7265512
6	7248529
11	7181644
15	7134565
20	7083173

```
ggplot(clg_ss) +
  geom_point(aes(x = x, y = y),alpha = 0.05) +
  geom_line(aes(x = x, y = prediction, color= df)) +
  facet_wrap(df ~ ., nrow = 2) +
  labs(title = "Smooth Spline")
```



The model obtained from CV method has the degree of freedom of 4.686 and lambda 0.031 has the lowest MSE in the model candidates. The fitted model is almost a smooth line. The MSE_{tr} is 7.266×10^6.

Q3

```
set.seed(123123)
cl = makePSOCKcluster(5)# if windows, set to 1

registerDoParallel(cl)

clg_gam =
  train(
    x = X_train,
    y = Y_train,
    method = "gam",
```

```r
    tuneGrid = expand.grid(select = c(T, F),
                           method  = c("GCV.cp", "REML")),
    metric = "RMSE",
    trControl = ctrl
  )

stopCluster(cl)

clg_gam$bestTune
```

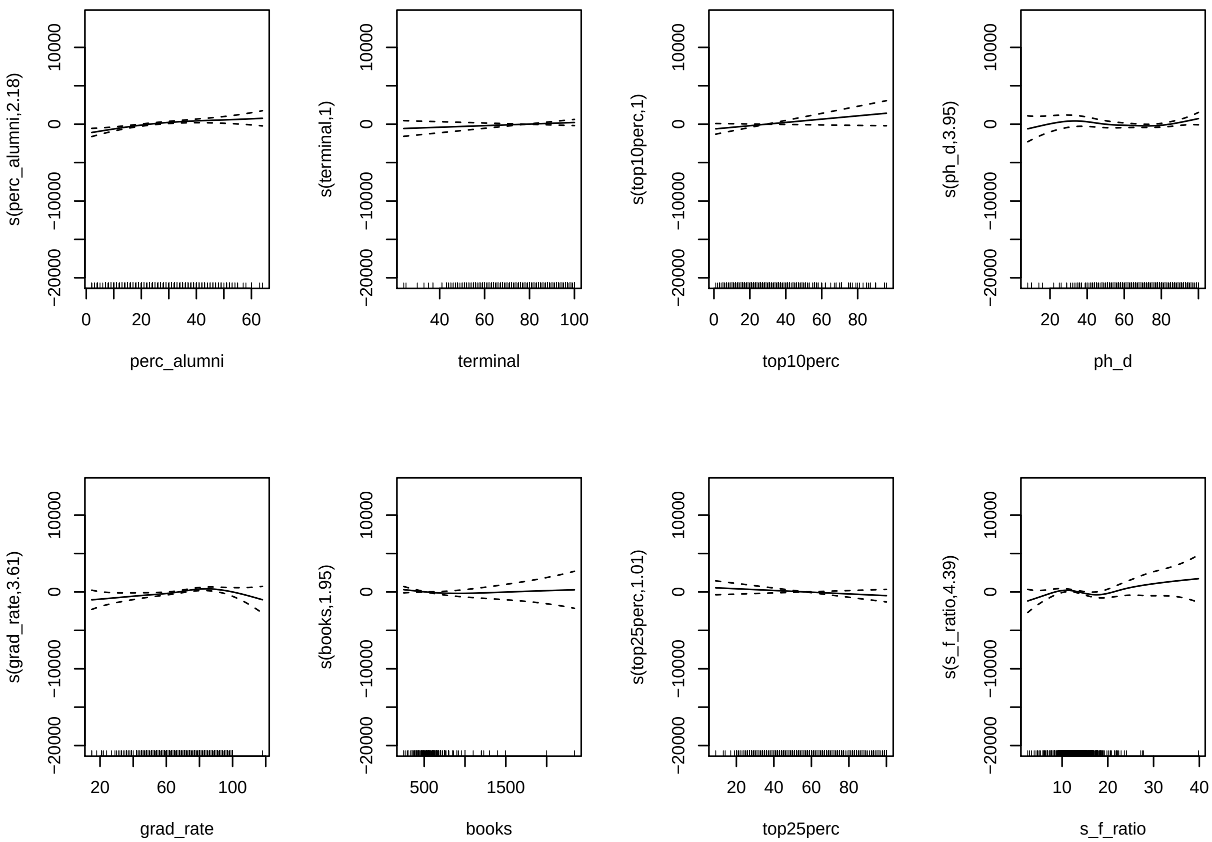
```
##   select method
## 2  FALSE   REML
```

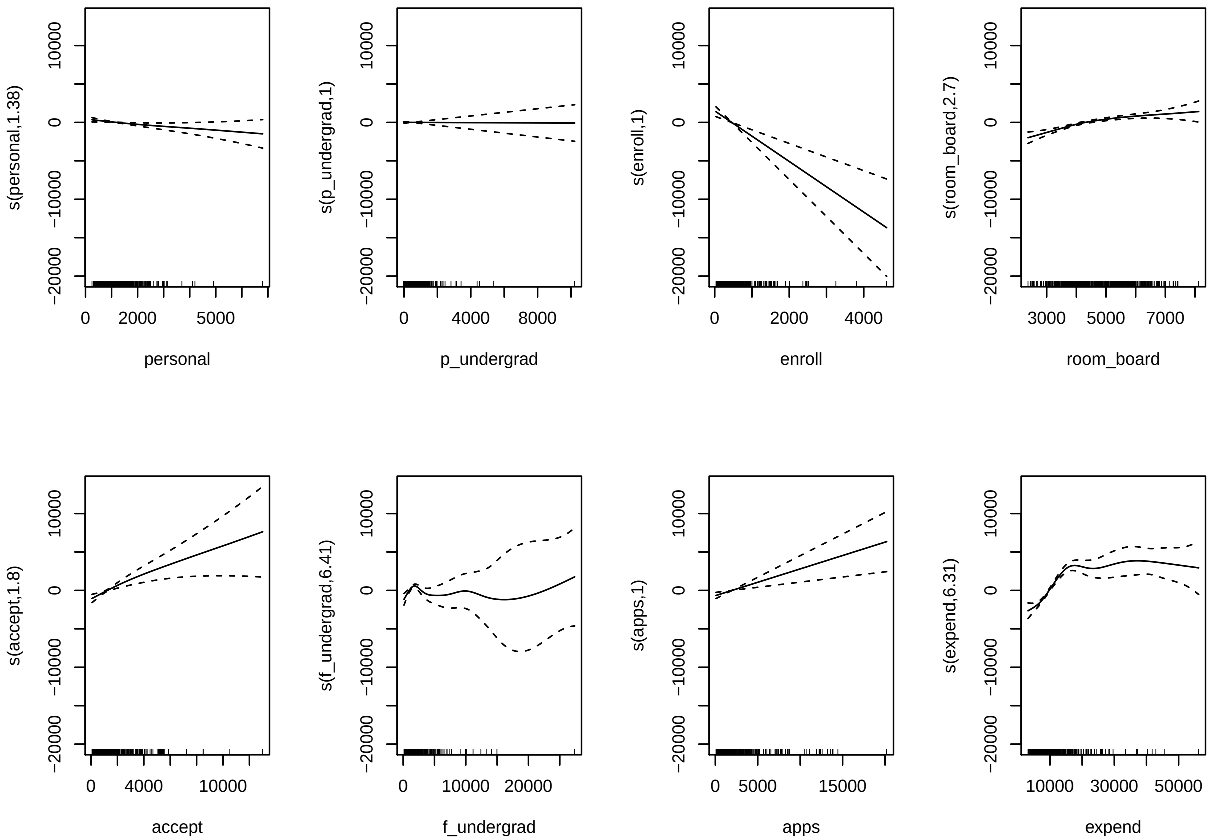
```r
clg_gam_mse = mean((Y_train-predict(clg_gam))^2)

summary(clg_gam$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)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11789.6       67.7     174   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df     F p-value
## s(perc_alumni) 2.18   2.77  8.28 7.5e-05 ***
## s(terminal)    1.00   1.00  1.21 0.27173
## s(top10perc)   1.00   1.00  3.05 0.08150 .
## s(ph_d)        3.95   4.91  1.93 0.07850 .
## s(grad_rate)   3.61   4.54  3.62 0.00536 **
## s(books)       1.95   2.44  1.25 0.38019
## s(top25perc)   1.01   1.02  1.38 0.23870
## s(s_f_ratio)   4.39   5.45  2.38 0.03602 *
## s(personal)    1.38   1.67  3.65 0.02507 *
## s(p_undergrad) 1.00   1.00  0.00 0.95512
## s(enroll)      1.00   1.01 18.63 1.8e-05 ***
## s(room_board)  2.70   3.44 16.78 < 2e-16 ***
## s(accept)      1.80   2.28  7.27 0.00051 ***
## s(f_undergrad) 6.41   7.43  4.45 7.2e-05 ***
## s(apps)        1.00   1.00 10.58 0.00122 **
## s(expend)      6.31   7.47 20.40 < 2e-16 ***
## ---
```

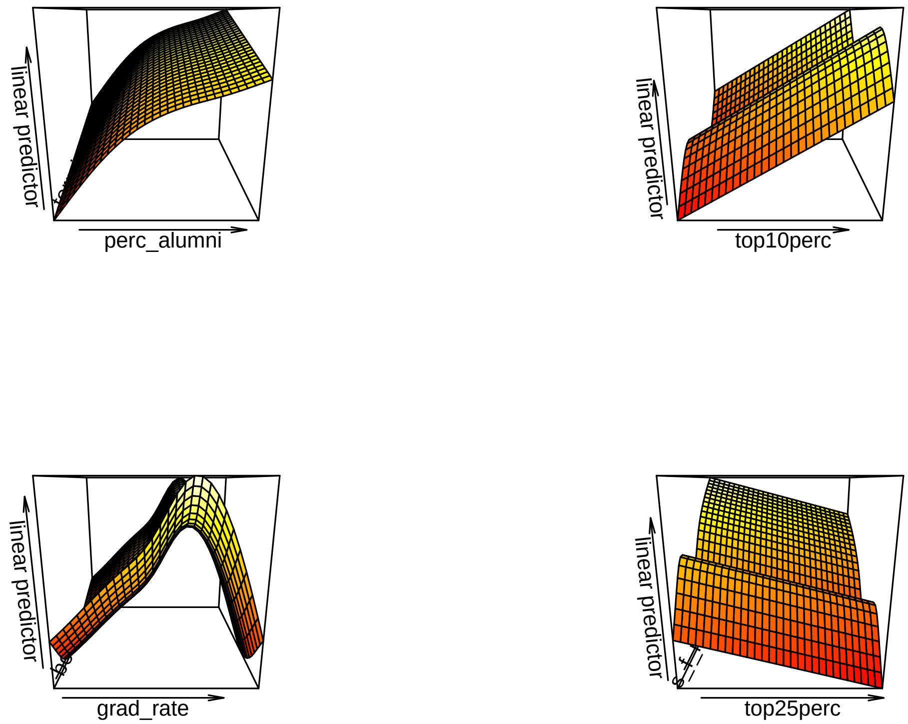
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.811   Deviance explained = 82.5%
## -REML = 4890.7  Scale est. = 2.5845e+06  n = 564
```

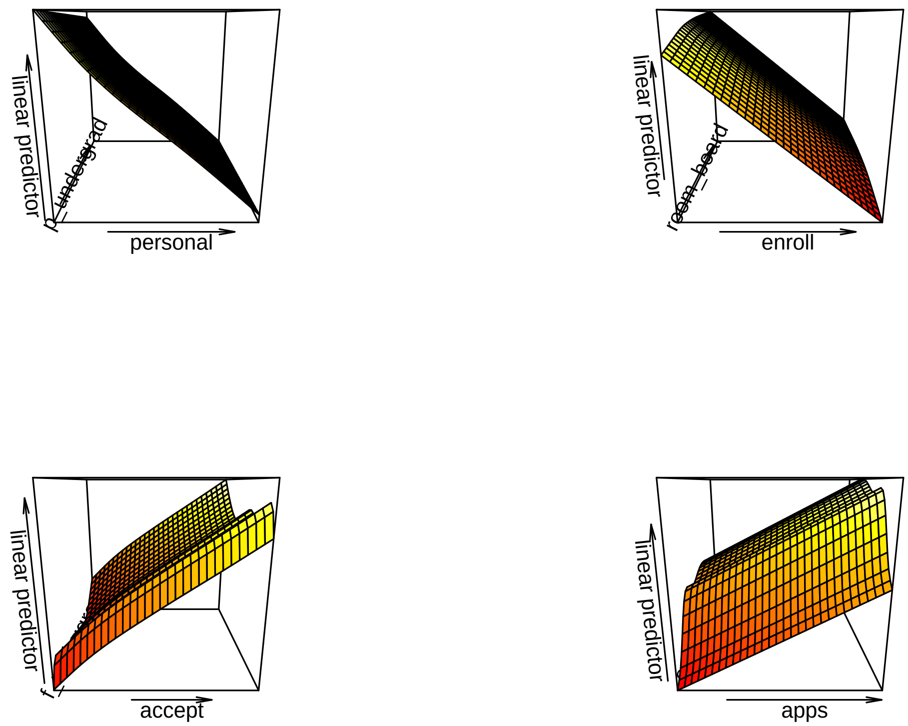
```r
par(mfrow = c(2,4))

plot(clg_gam$finalModel)
```






```r
par(mfrow=c(2,2))
for (i in 1:8){
  predictor = clg_gam$finalModel$terms %>% attr("term.labels") %>% .[(2*i-1):(2*i)]
  vis.gam(clg_gam$finalModel,predictor)
}
```


linear predictor
perc_alumni
linear predictor
top10perc
linear predictor
grad_rate
linear predictor
top25perc



Using caret tuning, the best tuning methods is `select = F` and `method = "REML"`. With this method, all variable is applied with spline function except for Indicator of `College` which is not selected by caret. The MSE_{tr} is 2.393×10^6.

Q4

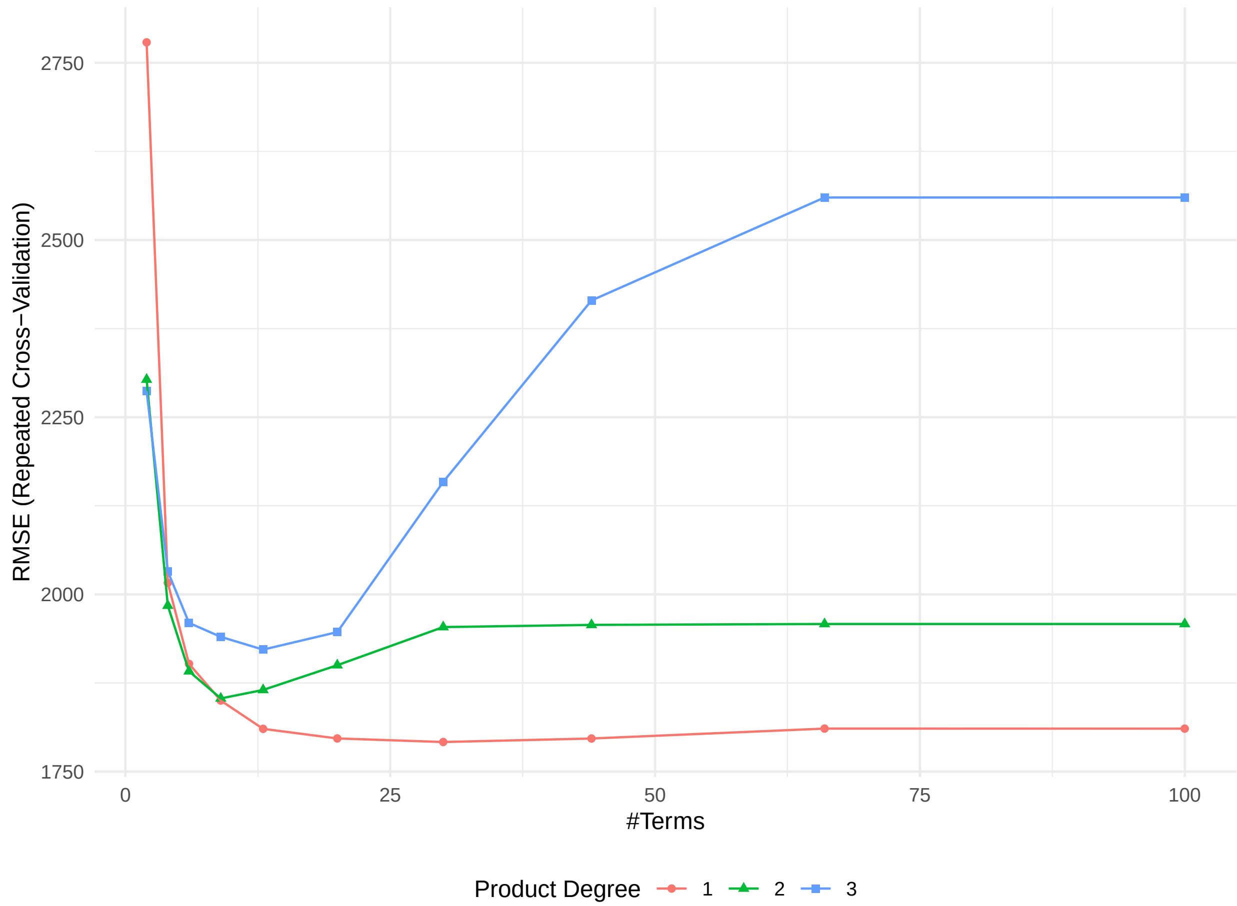
```
set.seed(123123)
cl = makePSOCKcluster(5) #if windows, set to 1
registerDoParallel(cl)
clg_mars =
  train(
    x = X_train,
    y = Y_train,
    method = "earth",
    tuneGrid = expand.grid(degree = 1:3,
                           nprune = exp(
                             seq(1, log(100), length = 10)
                           )%/%1),
    metric = "RMSE",
    trControl = ctrl
  )
```

```r
stopCluster(cl)

clg_mars$finalModel$coefficients %>%
  knitr::kable(caption = "Hints")
```

Table 5: Hints

	y
(Intercept)	10704.506
h(expend-15622)	-0.717
h(4440-room_board)	-1.234
h(grad_rate-95)	-166.256
h(95-grad_rate)	-26.247
h(f_undergrad-1350)	-0.339
h(1350-f_undergrad)	-1.396
h(21-perc_alumni)	-59.636
h(apps-3767)	0.347
h(1300-personal)	1.035
h(903-enroll)	3.934
h(2165-accept)	-1.867
collegeBennington College	6076.443
collegeWentworth Institute of Technology	-6358.623
collegeLivingstone College	-6012.630
collegeSpelman College	-5568.603
h(expend-5970)	0.738
collegeCreighton University	-6397.362
collegeTrinity University	-5915.358
collegeArkansas College (Lyon College)	-5548.820
collegeTuskegee University	-4692.058
collegeBuena Vista College	4389.802
collegeMorehouse College	-4289.216
collegeXavier University of Louisiana	-4376.861
collegeGreen Mountain College	4073.321
collegeWashington and Lee University	-3942.196
collegeHillsdale College	-3915.920
collegeBerry College	-4118.839
collegeWake Forest University	-4245.003
collegeSt. Paul's College	-3793.440

```r
ggplot(clg_mars)
```




```r
clg_mars$bestTune
```

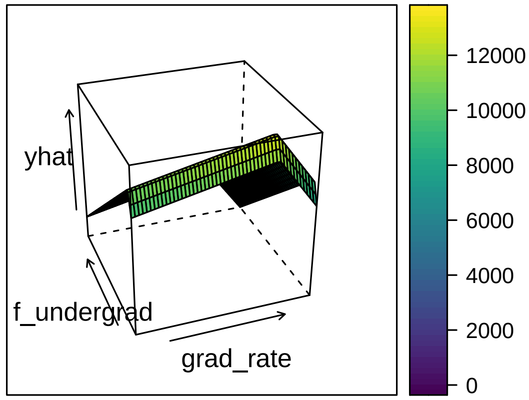
```
##   nprune degree
## 7     30      1
```

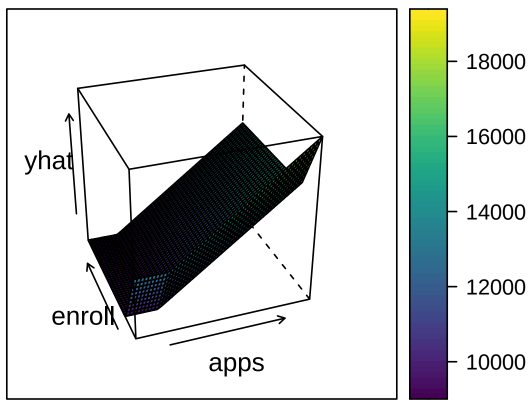
```r
summary(clg_mars$finalModel)
```

```
## Call: earth(x=matrix[564,580], y=c(7440,12280,11...), keepxy=TRUE, degree=1,
##             nprune=30)
##
##                                         coefficients
## (Intercept)                                    10705
## collegeArkansas College (Lyon College)         -5549
## collegeBennington College                       6076
## collegeBerry College                           -4119
## collegeBuena Vista College                      4390
## collegeCreighton University                    -6397
## collegeGreen Mountain College                   4073
## collegeHillsdale College                       -3916
## collegeLivingstone College                     -6013
## collegeMorehouse College                       -4289
## collegeSpelman College                         -5569
```

```
## collegeSt. Paul's College                        -3793
## collegeTrinity University                        -5915
## collegeTuskegee University                       -4692
## collegeWake Forest University                    -4245
## collegeWashington and Lee University             -3942
## collegeWentworth Institute of Technology         -6359
## collegeXavier University of Louisiana            -4377
## h(apps-3767)                                         0
## h(2165-accept)                                      -2
## h(903-enroll)                                        4
## h(1350-f_undergrad)                                 -1
## h(f_undergrad-1350)                                  0
## h(4440-room_board)                                  -1
## h(1300-personal)                                     1
## h(21-perc_alumni)                                  -60
## h(expend-5970)                                       1
## h(expend-15622)                                     -1
## h(95-grad_rate)                                    -26
## h(grad_rate-95)                                   -166
##
## Selected 30 of 69 terms, and 26 of 580 predictors (nprune=30)
## Termination condition: RSq changed by less than 0.001 at 69 terms
## Importance: expend, room_board, perc_alumni, accept, f_undergrad, apps, ...
## Number of terms at each degree of interaction: 1 29 (additive model)
## GCV 2336202    RSS 1.06e+09    GRSq 0.83    RSq 0.863
```

```r
p1 = pdp::partial(clg_mars, pred.var = c("grad_rate", "f_undergrad")) %>%
  plotPartial(
    levelplot = FALSE,
    zlab = "yhat",
    drape = TRUE,
    screen = list(z = 20, x = -60)
  )

p2 = pdp::partial(clg_mars, pred.var = c("apps", "enroll")) %>%
  plotPartial(
    levelplot = FALSE,
    zlab = "yhat",
    drape = TRUE,
    screen = list(z = 20, x = -60)
  )

grid.arrange(p1,p2,nrow = 2)
```





The final model has 3 degree and 30 hints in the model. total of 30 term and 26 predictors are includes in the model. The mse of the MARS model is 1.873×10^6

```
rmp = caret::resamples(list(gam = clg_gam,
                            mars = clg_mars))

summary(rmp)
```

```
##
## Call:
## summary.resamples(object = rmp)
##
## Models: gam, mars
## Number of resamples: 25
##
## MAE
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## gam  1170    1288   1351 1336    1374 1496    0
## mars 1223    1313   1379 1369    1421 1555    0
##
## RMSE
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## gam  1539    1649   1786 1757    1812 1991    0
## mars 1570    1696   1786 1792    1884 2034    0
```

```
##
## Rsquared
##       Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## gam  0.736   0.772  0.782 0.781   0.800 0.824    0
## mars 0.728   0.755  0.773 0.773   0.787 0.825    0
```