Homework 2

Jeffrey LIANG

2/20/2021

```
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

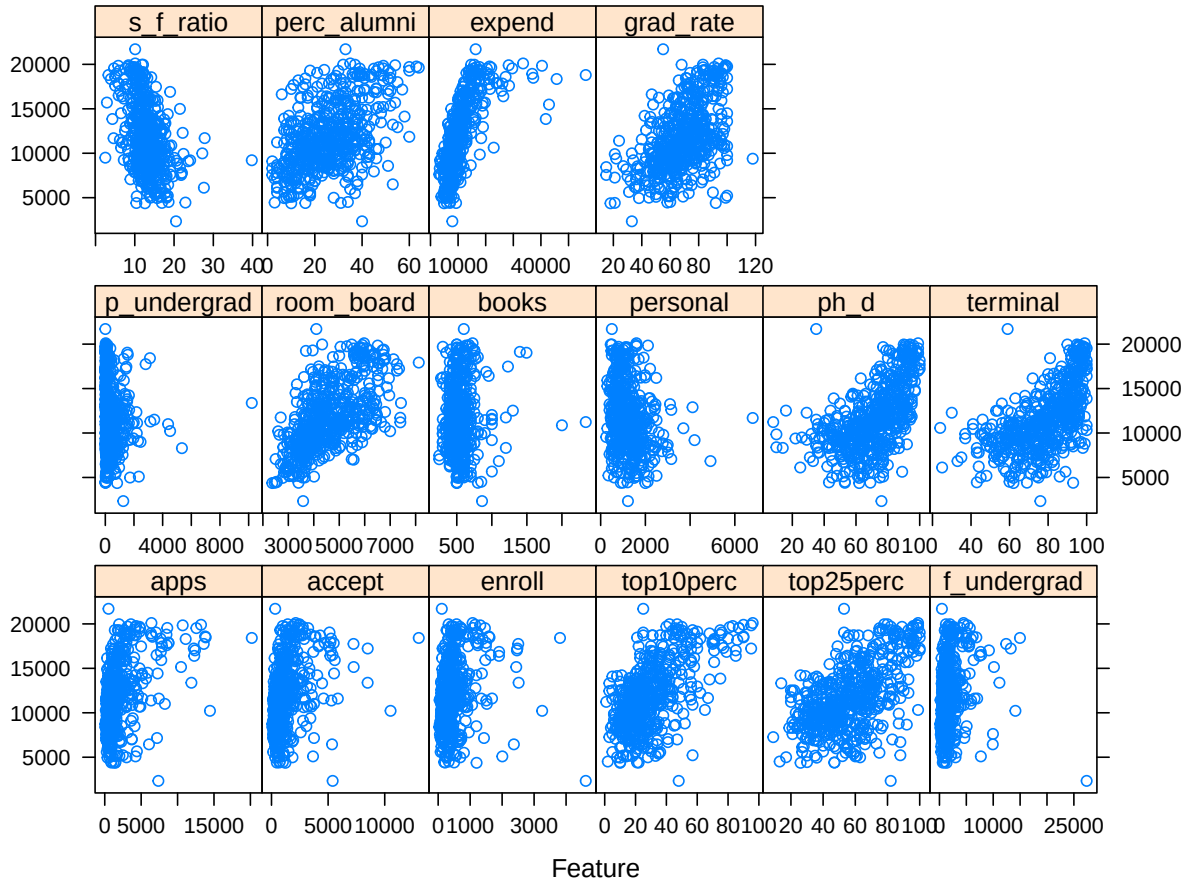
Variable type: numeric

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()

train_index = createDataPartition(clg_data$outstate, p = 0.8, list = F)

clg_train = clg_data[train_index,]

clg_test = clg_data[-train_index,]

Y_train = clg_train$outstate

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

Y_ts = clg_test$outstate

X_ts = model.matrix(outstate ~., data = clg_test)[,-1]

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

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



Q2

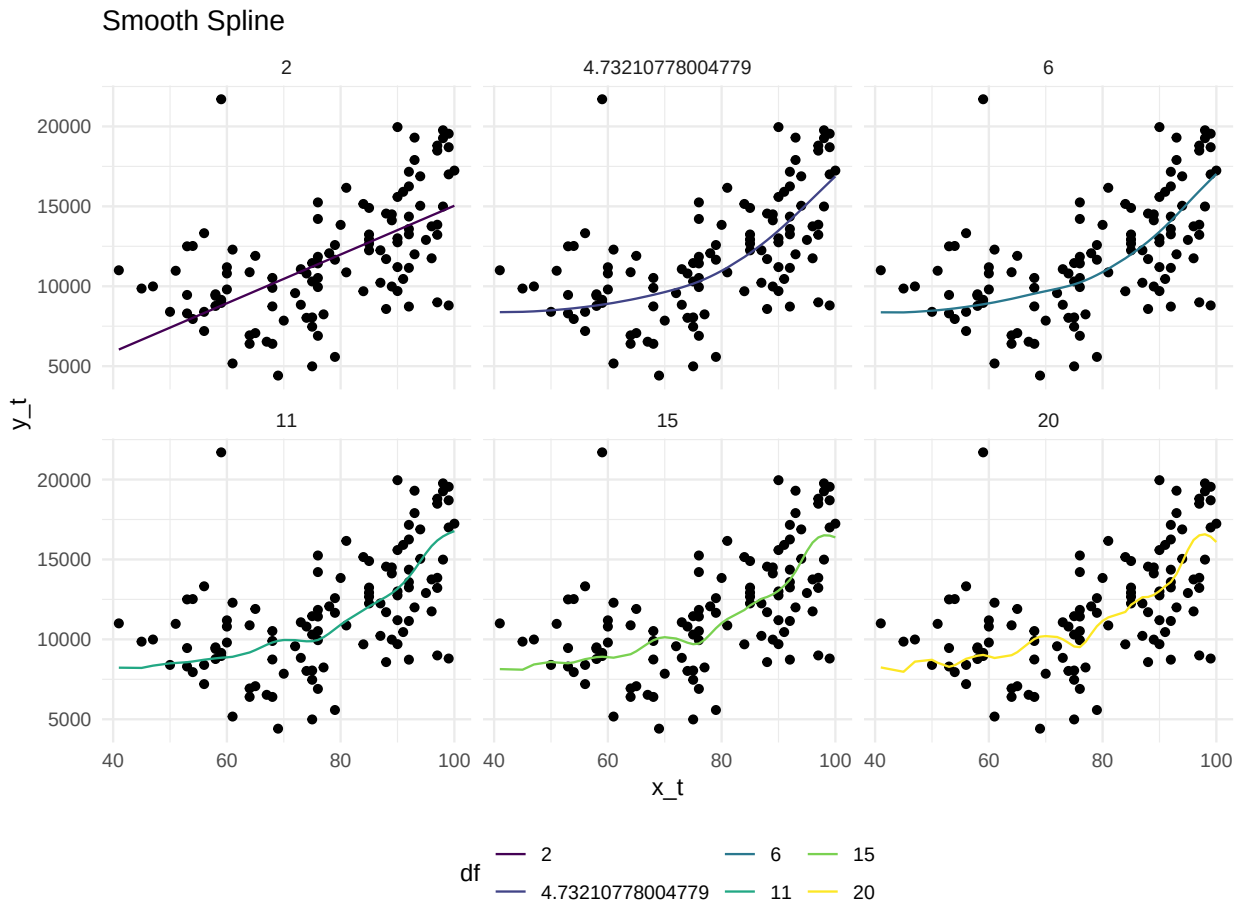
```
clg_ss_cv = smooth.spline(clg_train$terminal, Y_train, cv = T)

clg_ss =
  tibble(
    x = list(clg_train$terminal),
    y = list(Y_train),
    x_t = list(clg_test$terminal),
    y_t = list(Y_ts),
    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),
```

```
    x_t = list(clg_test$terminal),
    y_t = list(Y_ts),
    df = clg_ss_cv$df,
    model = list(clg_ss_cv)
  )) %>%
  mutate(
    prediction = map2(.x = x_t,
                      .y = model,
                      ~predict(object = .y,x = .x,se=F)$y),
    df = as.factor(df)
  ) %>%
  select(df, y_t, prediction, x_t) %>%
  unnest(c(prediction, y_t,x_t))

clg_ss %>%
  group_by(df) %>%
  summarise(mse =
              mean((y_t - 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	9722510
4.73210778004779	8621943
6	8658446
11	8770963
15	8866245
20	8965561

```
ggplot(clg_ss) +
  geom_point(aes(x = x_t, y = y_t)) +
  geom_line(aes(x = x_t, 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 4.732 and lambda 0.024 has the lowest MSE in the model candidates.

Q3

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

registerDoParallel(cl)

clg_gam =
  train(
    x = X_train,
    y = Y_train,
    method = "gam",
    tuneGrid = expand.grid(select = c(T, F),
                           method  = c("GCV.cp", "REML")),
    metric = "RMSE",
    trControl = ctrl
  )

stopCluster(cl)
```

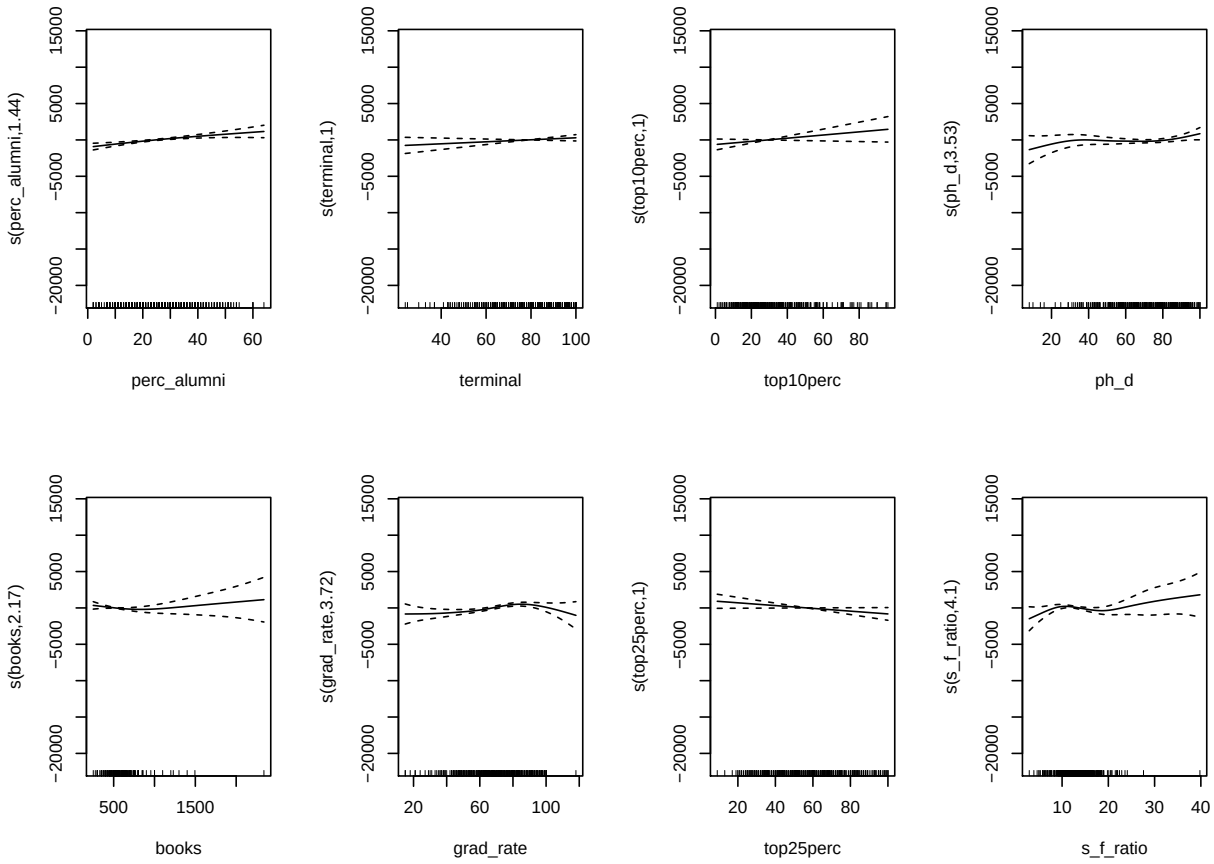
```
clg_gam$bestTune
```

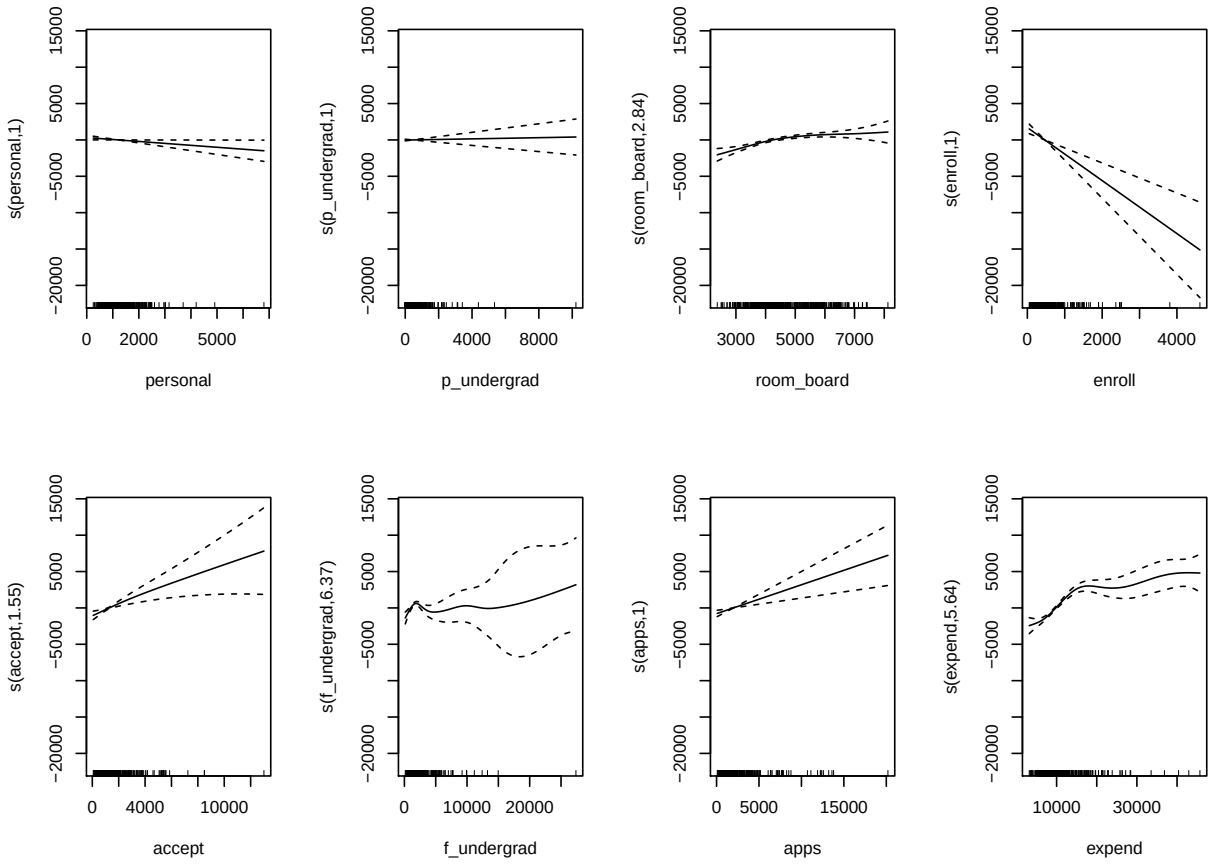
```
##   select method
## 2  FALSE   REML
```

```
summary(clg_gam$finalModel)
```

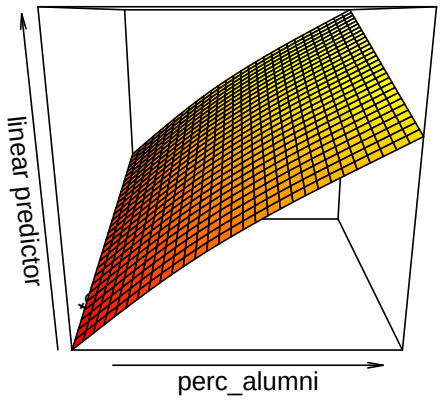
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(perc_alumni) + s(terminal) + s(top10perc) + s(ph_d) +
##     s(books) + s(grad_rate) + s(top25perc) + s(s_f_ratio) + s(personal) +
##     s(p_undergrad) + s(room_board) + s(enroll) + s(accept) +
##     s(f_undergrad) + s(apps) + s(expend)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11782.3       73.6     160   <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) 1.44   1.77  9.48 0.00018 ***
## s(terminal)    1.00   1.00  1.84 0.17584
## s(top10perc)   1.00   1.00  2.77 0.09706 .
## s(ph_d)        3.53   4.43  1.61 0.14692
## s(books)       2.17   2.73  1.49 0.30856
## s(grad_rate)   3.72   4.67  4.36 0.00121 **
## s(top25perc)   1.00   1.00  3.56 0.05996 .
## s(s_f_ratio)   4.10   5.08  2.02 0.07234 .
## s(personal)    1.00   1.00  4.16 0.04195 *
## s(p_undergrad) 1.00   1.00  0.10 0.74695
## s(room_board)  2.84   3.60 12.29 < 2e-16 ***
## s(enroll)      1.00   1.00 21.20 6.0e-06 ***
## s(accept)      1.55   1.89  6.21 0.00190 **
## s(f_undergrad) 6.37   7.41  4.70 3.8e-05 ***
## s(apps)        1.00   1.00 12.17 0.00054 ***
## s(expend)      5.64   6.83 17.82 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.822   Deviance explained = 83.7%
## -REML = 3888.2  Scale est. = 2.4452e+06  n = 452
```

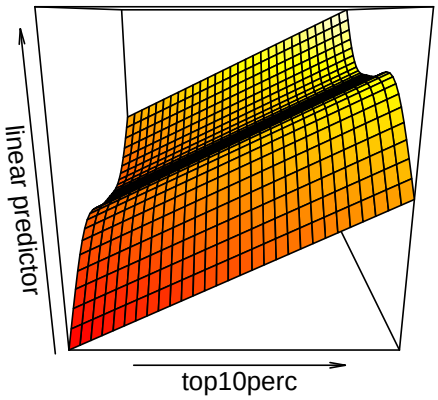
```
par(mfrow = c(2,4))

plot(clg_gam$finalModel)
```

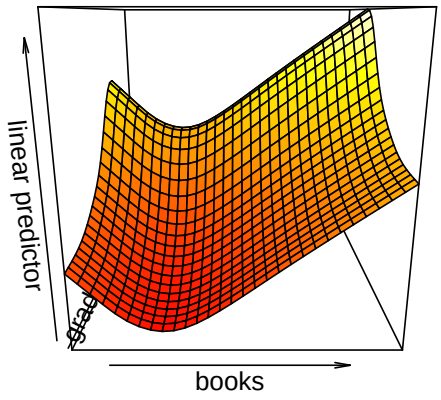

s(perc_alumni,1.44)
perc_alumni
s(terminal,1)
terminal
s(top10perc,1)
top10perc
s(ph_d,3.53)
ph_d
s(books,2.17)
books
s(grad_rate,3.72)
grad_rate
s(top25perc,1)
top25perc
s(s_f_ratio,4.1)
s_f_ratio

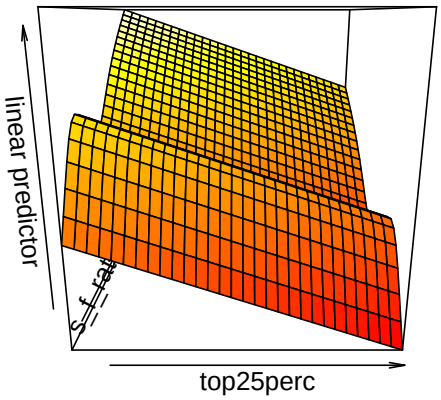


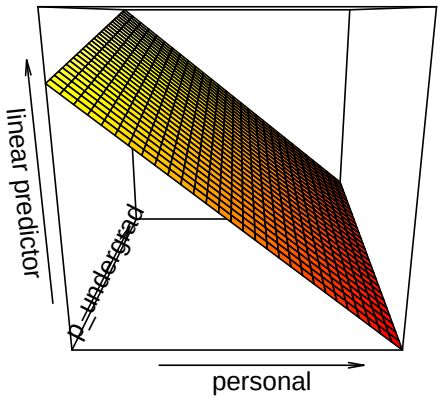

```
par(mfrow=c(1,2))
for (i in 1:8){
  predictor = clg_gam$finalModel$terms %>% attr("term.labels") %>% .[(2*i-1):(2*i)]
  vis.gam(clg_gam$finalModel,predictor)
}
```

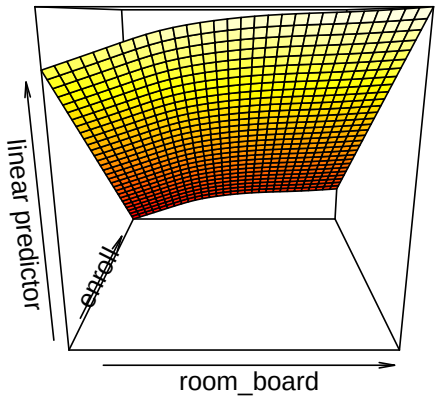

linear predictor
perc_alumni

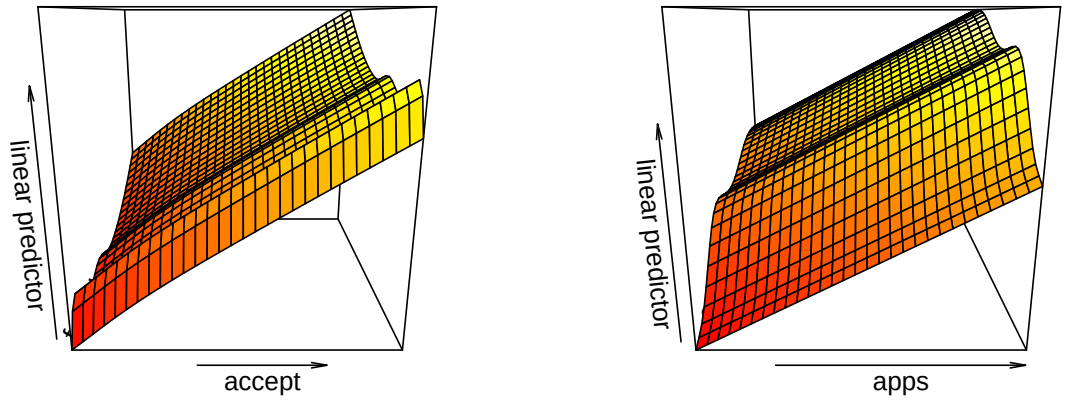

linear predictor
top10perc


linear predictor
books


linear predictor
top25perc


linear predictor
personal
personal_0


linear predictor
room_board
room_board



Using caret tuning, the best tuning methods is `select = F` and `method = "REML"`

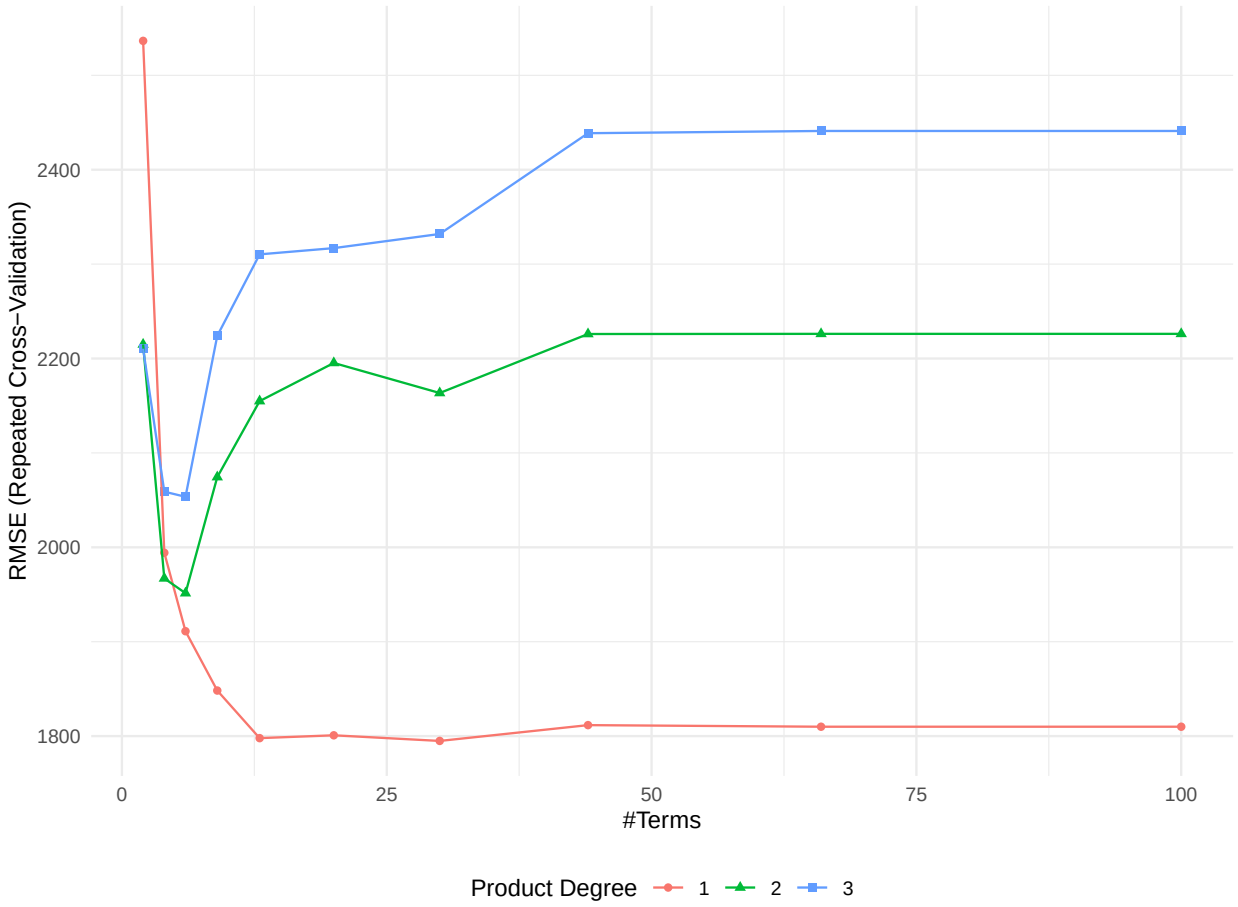
Q4

```
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
  )
stopCluster(cl)

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

Table 5: Hints

	y
(Intercept)	9607.674
h(expend-15494)	-0.660
h(grad_rate-88)	-86.144
h(88-grad_rate)	-33.388
h(4440-room_board)	-1.373
h(1442-f_undergrad)	-1.184
h(22-perc_alumni)	-63.977
h(apps-1422)	0.514
h(enroll-913)	-2.212
h(913-enroll)	5.000
h(1579-accept)	-1.934
collegeSpelman College	-6755.687
collegeCreighton University	-5459.568
h(expend-6869)	0.739
collegeWentworth Institute of Technology	-6637.781
h(8.8-s_f_ratio)	-552.582
collegeLivingstone College	-5508.517
collegeTrinity University	-5281.317
collegeArkansas College (Lyon College)	-5450.354
collegeTuskegee University	-4613.500
collegeMorehouse College	-5206.570
h(1345-personal)	0.708
collegeAlbertson College	3885.203
collegeGreen Mountain College	3908.495
collegeXavier University of Louisiana	-3859.007
collegeBerry College	-4105.535
collegeChapman University	3683.263
collegeHillsdale College	-3755.306
collegeLouisiana College	-3522.610
collegeGrinnell College	-3300.535

```
ggplot(clg_mars)
```




```
clg_mars$bestTune
```

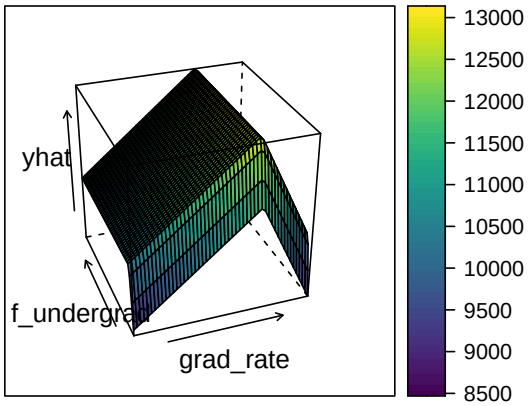
```
##   nprune degree
## 7     30      1
```

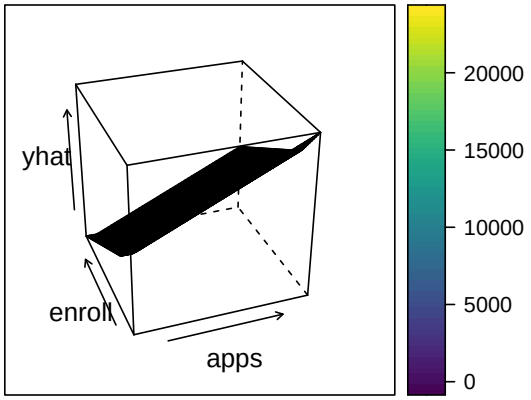
```
summary(clg_mars$finalModel)
```

```
## Call: earth(x=matrix[452,580], y=c(7440,12280,11...), keepxy=TRUE, degree=1,
##             nprune=30)
##
##                                          coefficients
## (Intercept)                                      9608
## collegeAlbertson College                         3885
## collegeArkansas College (Lyon College)          -5450
## collegeBerry College                            -4106
## collegeChapman University                        3683
## collegeCreighton University                     -5460
## collegeGreen Mountain College                    3908
## collegeGrinnell College                         -3301
## collegeHillsdale College                        -3755
## collegeLivingstone College                      -5509
## collegeLouisiana College                        -3523
```

```
## collegeMorehouse College                          -5207
## collegeSpelman College                            -6756
## collegeTrinity University                         -5281
## collegeTuskegee University                        -4613
## collegeWentworth Institute of Technology          -6638
## collegeXavier University of Louisiana             -3859
## h(apps-1422)                                          1
## h(1579-accept)                                       -2
## h(913-enroll)                                         5
## h(enroll-913)                                        -2
## h(1442-f_undergrad)                                  -1
## h(4440-room_board)                                   -1
## h(1345-personal)                                      1
## h(8.8-s_f_ratio)                                   -553
## h(22-perc_alumni)                                   -64
## h(expend-6869)                                        1
## h(expend-15494)                                      -1
## h(88-grad_rate)                                      -33
## h(grad_rate-88)                                      -86
##
## Selected 30 of 76 terms, and 26 of 580 predictors (nprune=30)
## Termination condition: RSq changed by less than 0.001 at 76 terms
## Importance: expend, grad_rate, room_board, accept, enroll, perc_alumni, ...
## Number of terms at each degree of interaction: 1 29 (additive model)
## GCV 2197690    RSS 7.51e+08    GRSq 0.84    RSq 0.879
```

```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)
```






```
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  1185    1255   1298 1328    1359 1584    0
## mars 1150    1325   1363 1380    1434 1609    0
##
## RMSE
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## gam  1521    1613   1690 1741    1762 2361    0
## mars 1503    1679   1787 1795    1912 2101    0
##
## Rsquared
```

```
##       Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## gam  0.684   0.770  0.796 0.783    0.81 0.845    0
## mars 0.704   0.758  0.773 0.771    0.79 0.828    0
```