Homework 2

Jeffrey LIANG

2/20/2021

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

Q1

Table 1: Data summary

Name	clg_data
Number of rows	564
Number of columns	18
Column type frequency:	
factor	1
numeric	17
Group variables	None

Variable type: factor

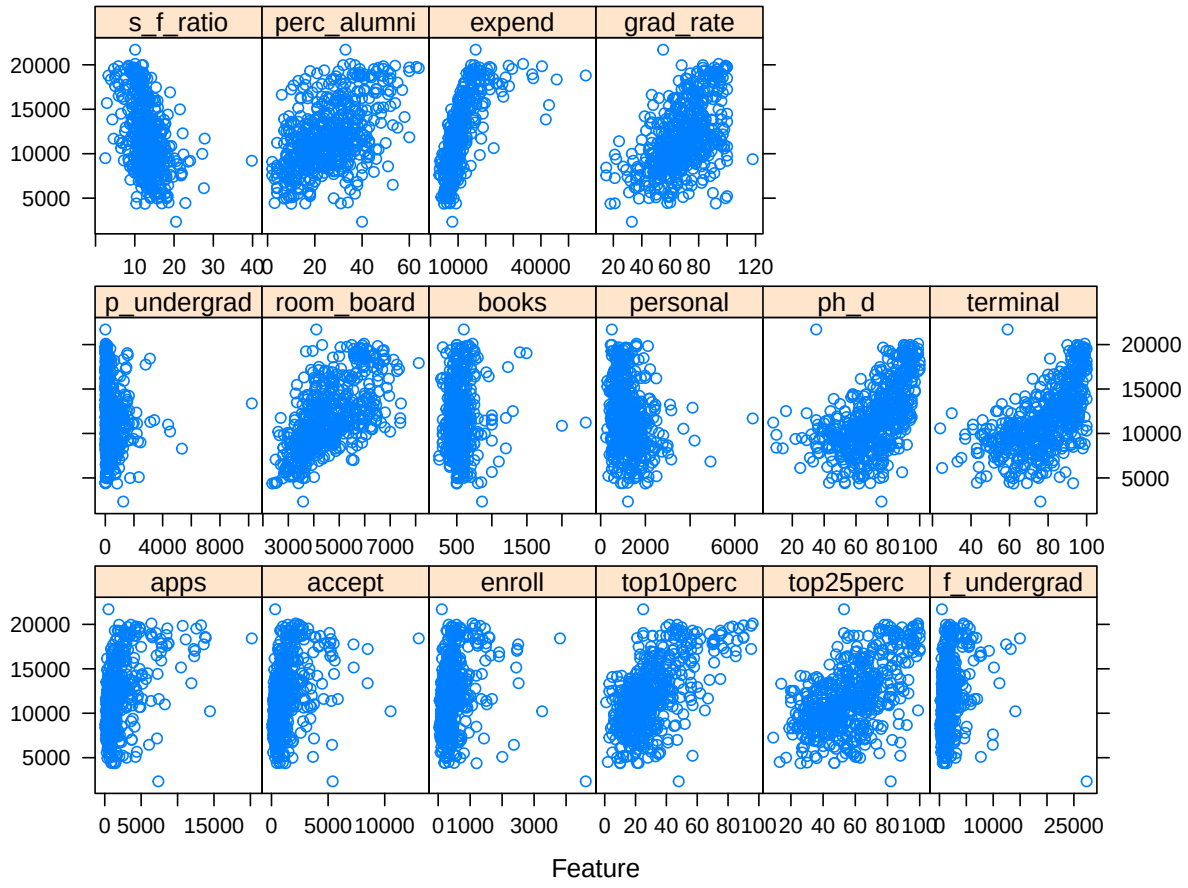
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
college	0	1	FALSE	564	Abi: 1, Ade: 1, Adr: 1, Agn: 1

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
apps	0	1	1969.5	2437.19	81.0	618.2	1132.5	2177.0	20192.0
accept	0	1	1304.6	1370.51	72.0	501.0	859.0	1579.2	13007.0
enroll	0	1	456.2	457.60	35.0	205.5	328.0	515.5	4615.0
top10perc	0	1	29.2	17.75	1.0	16.8	25.0	36.0	96.0
top25perc	0	1	56.9	19.54	9.0	42.0	55.0	70.0	100.0
f_undergrad	0	1	1869.5	2111.58	139.0	840.0	1272.5	1995.5	27378.0
p_undergrad	0	1	434.6	722.83	1.0	63.0	207.5	541.0	10221.0
outstate	0	1	11789.6	3699.59	2340.0	9100.0	11200.0	13962.5	21700.0
room_board	0	1	4582.5	1087.14	2370.0	3735.8	4400.0	5400.0	8124.0
books	0	1	547.5	175.09	250.0	450.0	500.0	600.0	2340.0
personal	0	1	1216.1	632.27	250.0	800.0	1100.0	1500.0	6800.0
ph_d	0	1	71.0	17.33	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	15.44	24.0	68.0	81.0	92.0	100.0
s_f_ratio	0	1	13.0	3.51	2.5	11.1	12.7	14.5	39.8
perc_alumni	0	1	25.9	12.41	2.0	16.0	25.0	34.0	64.0
expend	0	1	10450.6	5623.73	3186.0	7468.8	8954.0	11577.0	56233.0
grad_rate	0	1	69.0	16.72	15.0	58.0	69.0	81.0	118.0

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

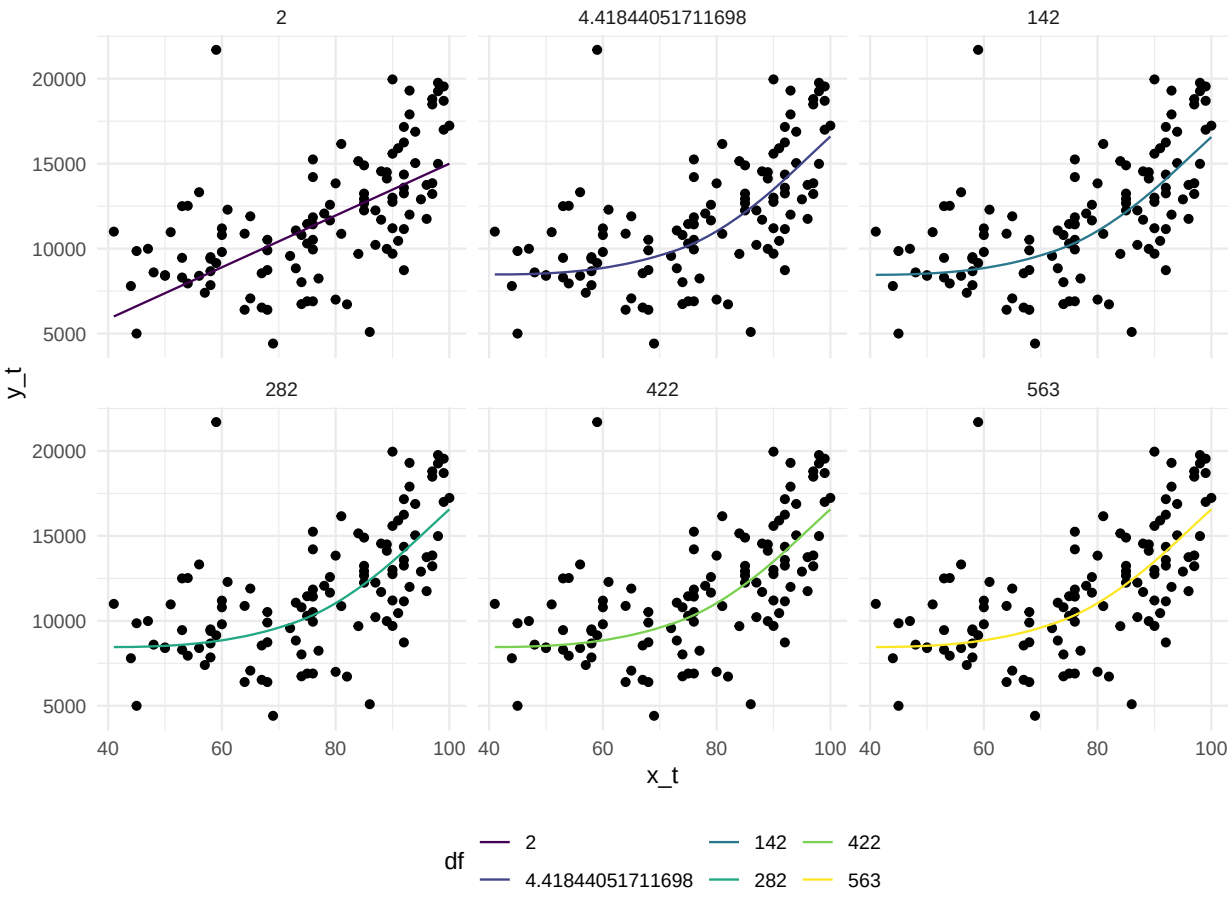


Q2

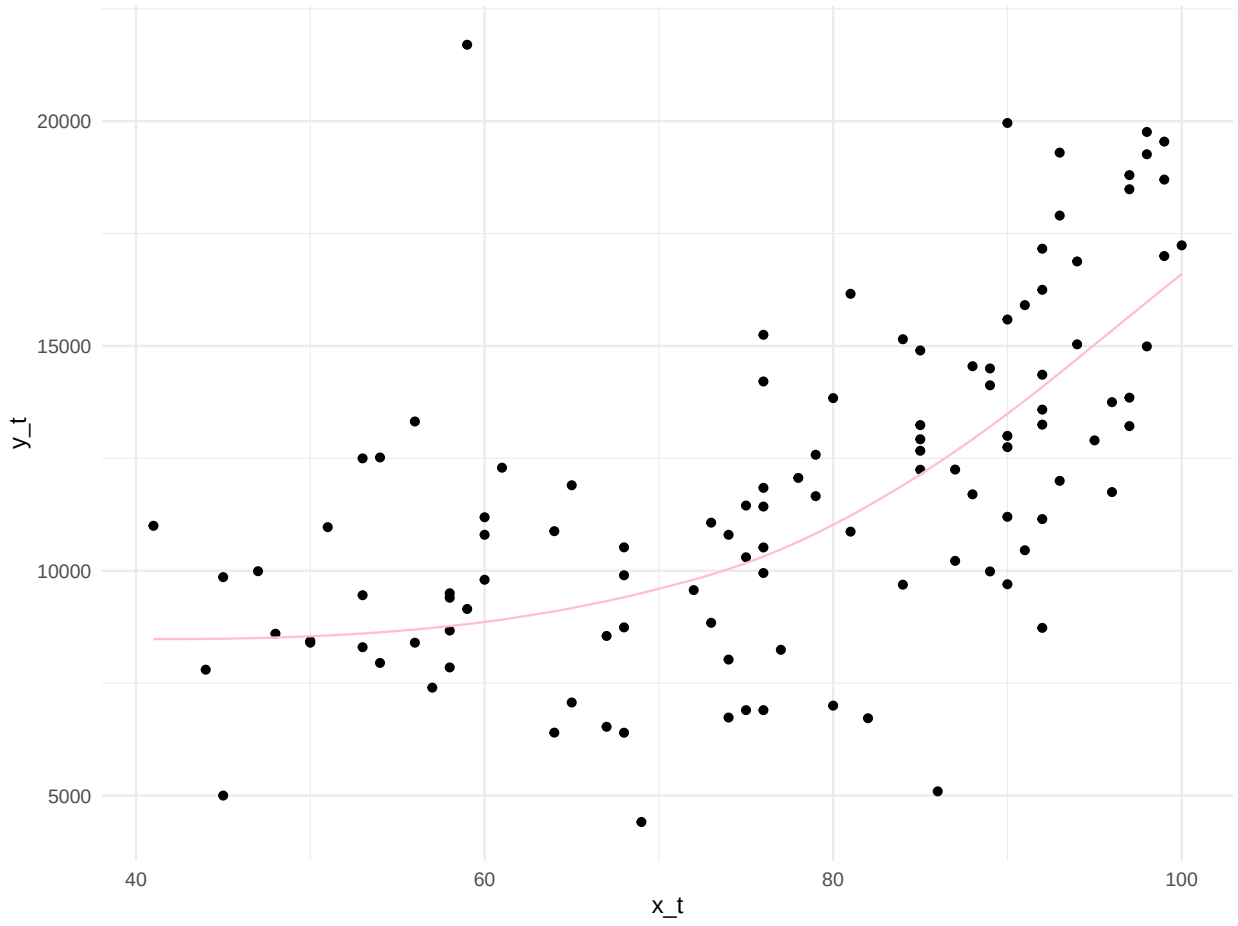
```
clg_ss_cv = smooth.spline(clg_train$terminal, Y_train, cv = T)
clg_ss =
  tibble(
    x = list(clg_train$terminal),
    y = list(Y_train),
    x_t = list(clg_test$terminal),
```

```
    y_t = list(Y_ts),
    df = list(seq(2, (nrow(
      clg_data
    ) - 1), 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))

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)
```




```
clg_ss %>%
  filter(df == clg_ss_cv$df) %>%
  arrange(x_t) %>%
  ggplot()+
  geom_point(aes(x_t,y_t))+
  geom_line(aes(x_t,prediction),color = "pink")
```

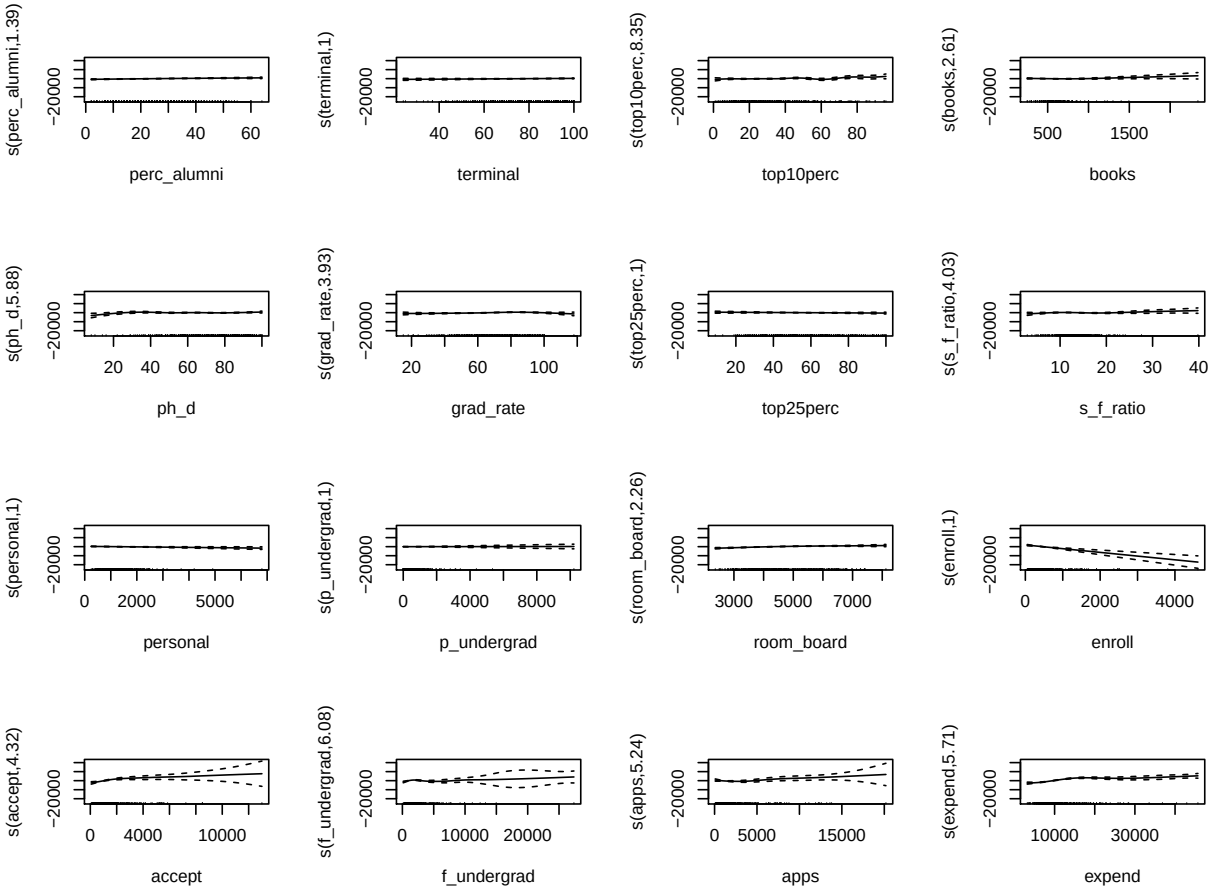


Q3

```
cl = makePSOCKcluster(5)# if windows, set to 1
registerDoParallel(cl)
clg_gam =
  train(x = X_train,
        y = Y_train,
        method = "gam",
        trControl = ctrl
        )
stopCluster(cl)

par(mfrow = c(4,4))

plot(clg_gam$finalModel)
```




```
vis.gam(clg_gam$finalModel,c("expend","apps"))
```

Q4

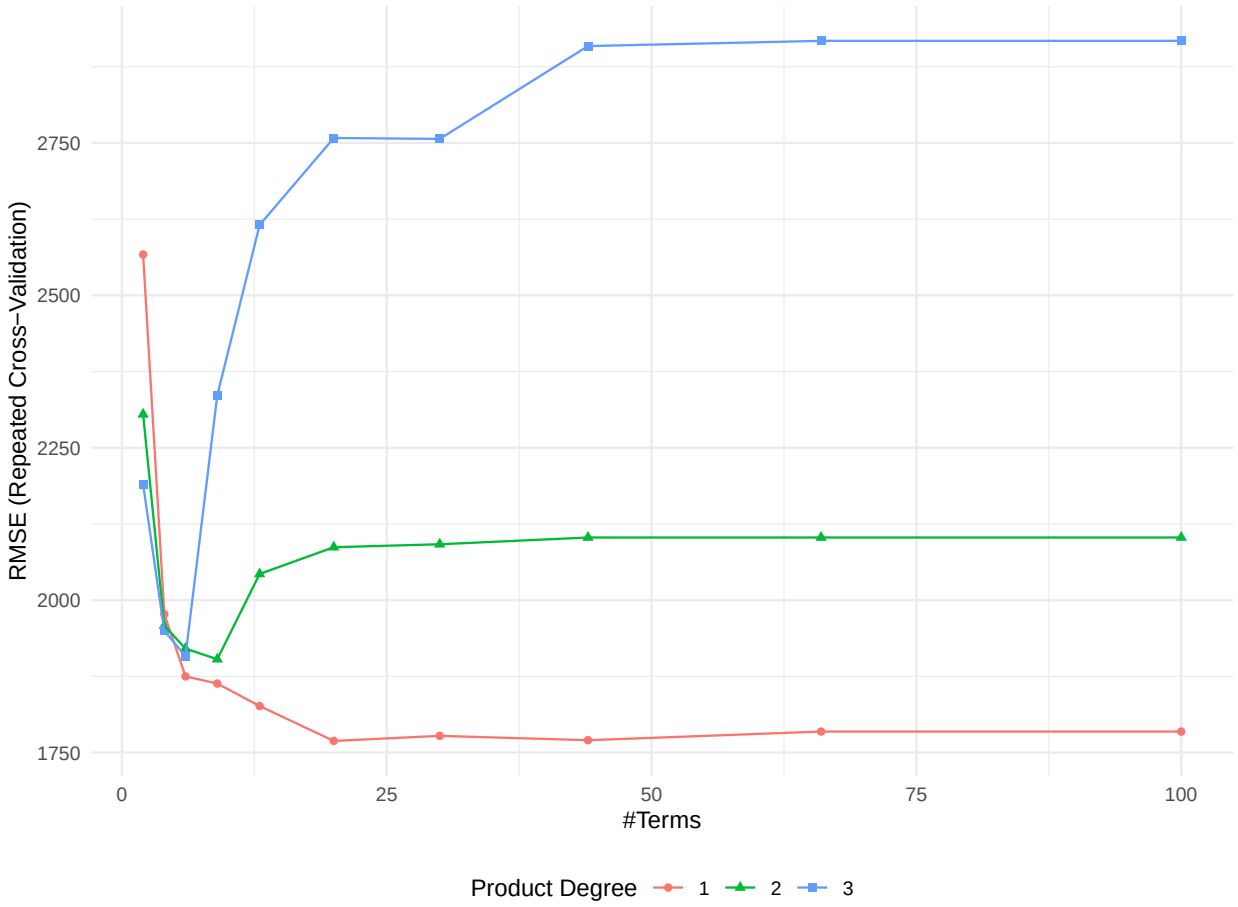
```
cl = makePSOCKcluster(5) #if windows, set to 1
registerDoParallel(cl)
clg_mars =
  train(
    x = X_train,
    y = Y_train,
    method = "earth",
    tuneGrid = expand.grid(degree = 1:3,
                           nprune = exp(
                             seq(1, log(100), length = 10)
         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

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

Table 4: Hints

	y
(Intercept)	10674.568
h(expend-14820)	-0.629
h(88-grad_rate)	-33.819
h(4328-room_board)	-1.540
h(f_undergrad-1427)	-0.936
h(apps-1416)	0.598
h(enroll-973)	-3.694
h(1553-accept)	-1.388
collegeCreighton University	-5384.899
collegeWentworth Institute of Technology	-5683.415
h(1300-personal)	1.165
collegeMorehouse College	-6187.455
collegeLivingstone College	-6525.707
collegeTrinity University	-5999.564
collegeArkansas College (Lyon College)	-5714.304
h(expend-5531)	0.726
h(8.8-s_f_ratio)	-485.876
collegeBuena Vista College	4466.642
h(f_undergrad-4540)	1.223
h(grad_rate-96)	-257.141

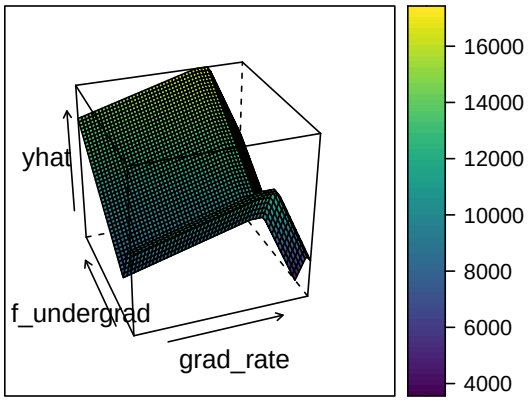
```
ggplot(clg_mars)
```

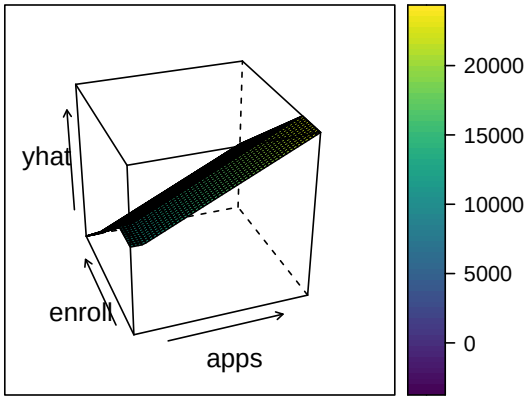



```
p1 = pdp::partial(clg_mars, pred.var = c("grad_rate", "f_undergrad")) %>%
  plotPartial(
    levelplot = FALSE,
    zlab = "yhat",
    drape = TRUE,
    screen = list(z = 20, x = -60)
  )

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)
```


yhat
f_undergrad
grad_rate
16000
14000
12000
10000
8000
6000
4000


yhat
enroll
apps
20000
15000
10000
5000
0