hw2

Amanda Howarth
3/19/2020

```
library(caret)
library(splines)
library(dplyr)
library(mgcv)
library(tidyverse)
library(ggplot2)
library(tidyverse)
library(readxl)
library(pdp)
library(earth)
```

LOADING DATA

```
colleges=
 read_excel(path = "./data/college.xlsx", sheet = 1) %>%
 janitor::clean_names() %>%
 filter(college != "Columbia University") %>%
 select(-college, apps, accept, enroll, top10perc, top25perc, f_undergrad, p_undergrad, room_board,
books, personal, ph_d, terminal, s_f_ratio, perc_alumni, expend, grad_rate, outstate) %>%
 na.omit()
```

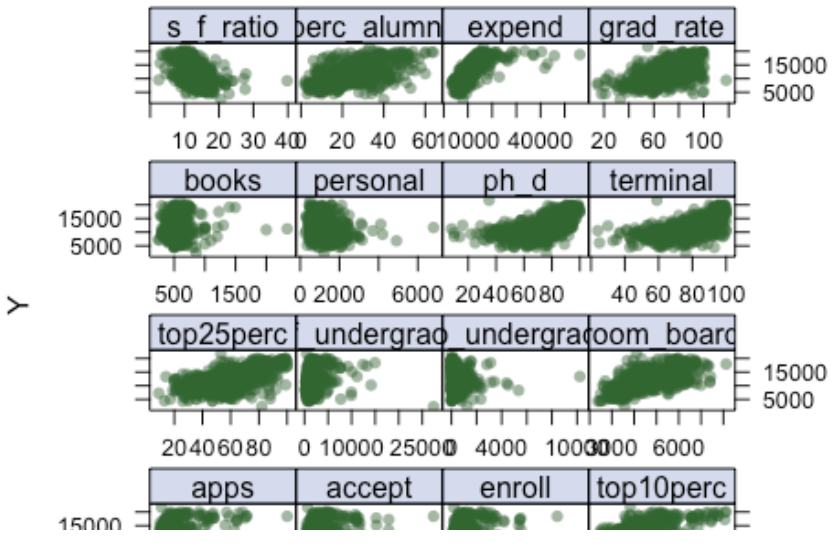
QUSETION 1A

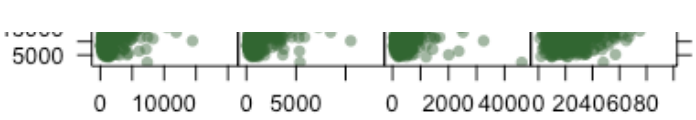
Create scatter plots of response vs. predictors.

```
# matrix of predictors
x <- model.matrix(outstate~.,colleges)[,-1]

# vector of response
y <- colleges$outstate
```

SCATTERPLOT

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



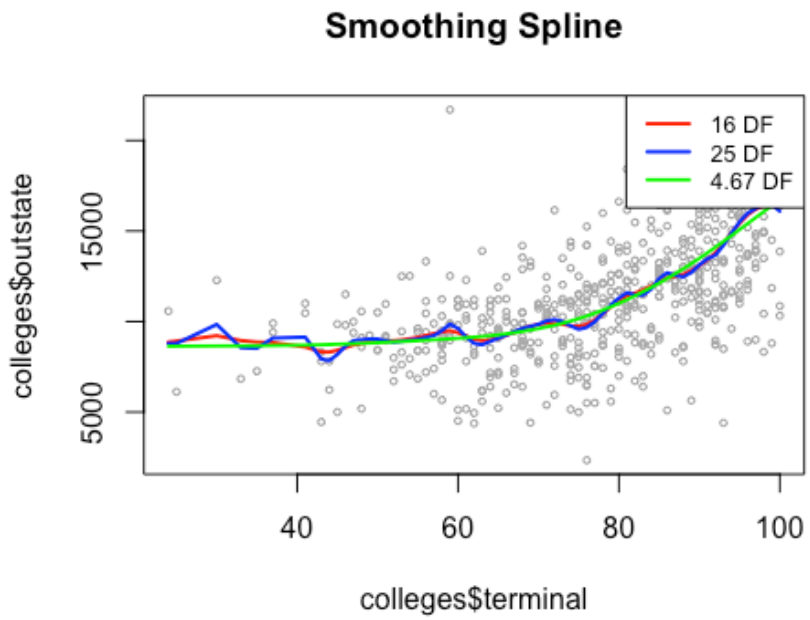


Sixteen scatter plots are presented - each displaying the relationship between one predictor and the outcome (out of state tuition).

QUESTION 1B

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.

```
terminallims <- range(colleges$terminal)
terminal.grid <- seq(from = terminallims[1],to = terminallims[2])

plot(colleges$terminal, colleges$outstate, xlim=terminallims, cex = .5, col = "darkgrey")
title ("Smoothing Spline")

fit1 = smooth.spline(colleges$terminal, colleges$outstate, df = 16)
fit2 = smooth.spline(colleges$terminal, colleges$outstate, df = 25)
fit3 = smooth.spline(colleges$terminal, colleges$outstate)
fit3$df
## [1] 4.468629
pred.ss <- predict(fit1, x = terminal.grid)
pred.ss <- predict(fit2, x = terminal.grid)
pred.ss <- predict(fit3, x = terminal.grid)

lines(fit1, col ="red", lwd =2)
lines(fit2, col = "blue", lwd = 2)
lines(fit3, col = "green", lwd = 2)
legend ("topright",legend =c("16 DF " ,"25 DF", "4.67 DF"), col=c("red "," blue ", "green"),
       lty =1, lwd =2, cex =.8)
```

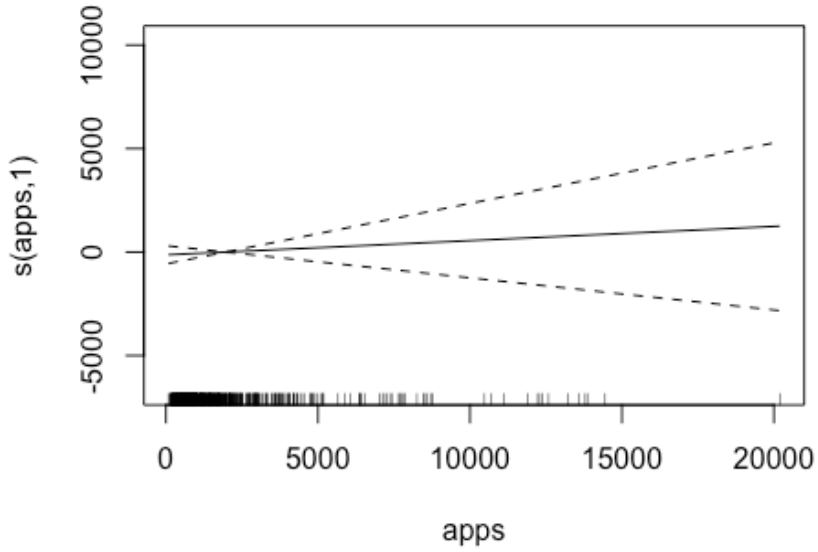


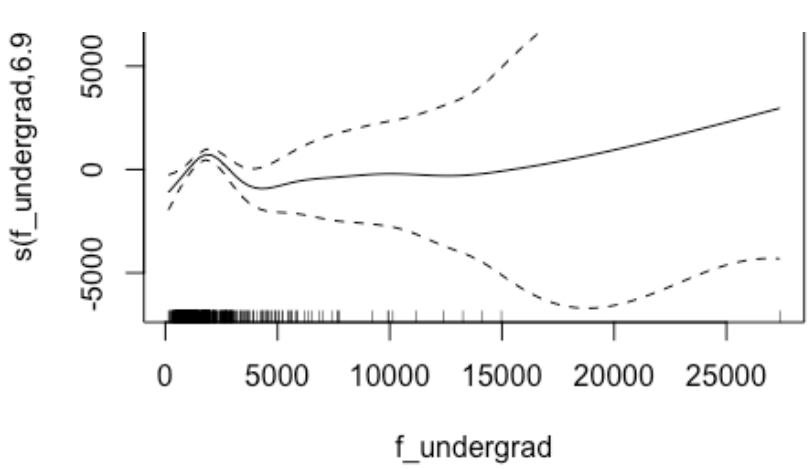
In the function “fit1” (red line), we specified df=16, and in the function “fit2” (blue line), we specified df= 25. The first function (fit1) determined which value of λ would lead to 16 degrees of freedom, and the second function (fit2) determined which value of λ would lead to 25 degrees of freedom. In the third function “fit 3” (green line), the smoothness level was chosen by generalized cross validation, which resulted in a value of λ that yields 4.69 degrees of freedom. We can see from the plot of the smoothing splines, that the green line (4.67 DF) is the smoothest line, while the other two lines are much more wiggly and overfit the data. The blue line (25 DF) overfits more than the red line (16 DF). It is recommended to use the smooth spline model whose degrees of freedom were chosen by cross validation.

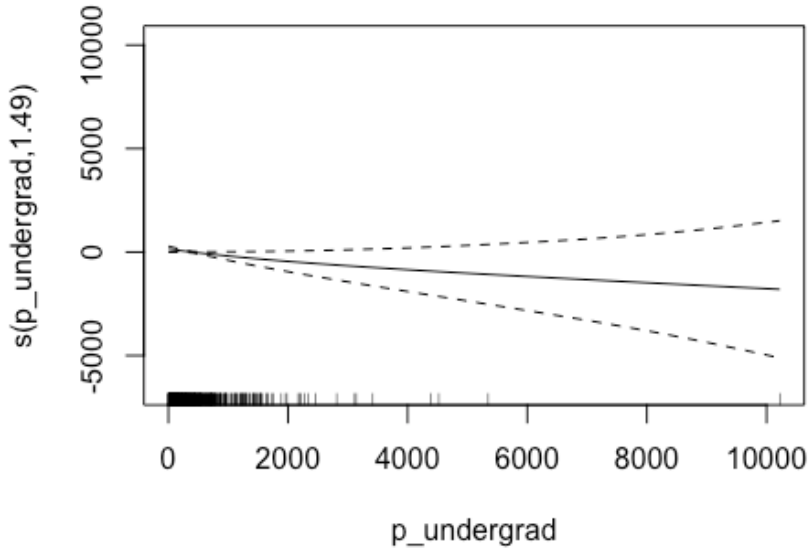
QUESTION 1C

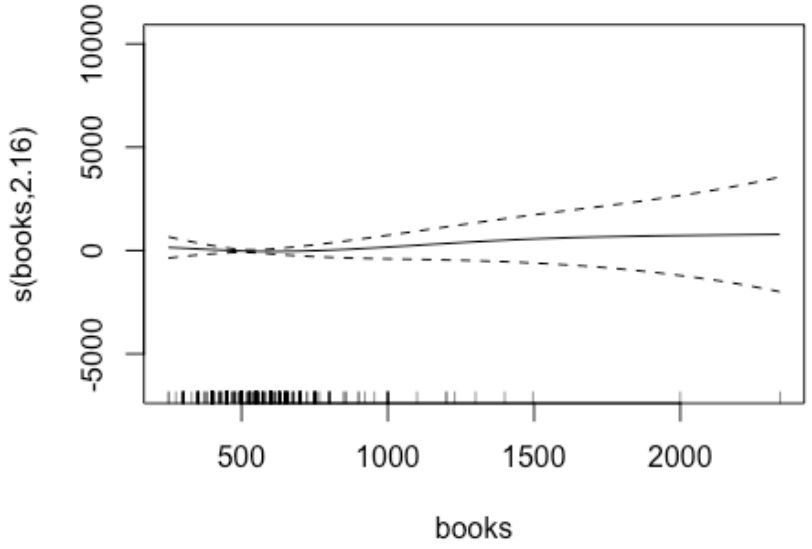
Fit a generalized additive model (GAM) using all the predictors. Plot the results and explain your findings.

```
#simple linear model. All terms treated as linear terms.
gam.m1 <- gam(outstate~
apps+accept+enroll+top10perc+top25perc+f_undergrad+p_undergrad+room_board+books+person
al+ph_d+terminal+s_f_ratio+perc_alumni+expend+grad_rate, data = colleges)

#After looking at the scatterplots in Question 1A, we think there may be non-linear trends between
some predictors (such as f_undegrad, p_undergrad, books, and apps) and the response variable,
and thus, we added s functions.
gam.m2 <- gam(outstate~
s(apps)+accept+enroll+top10perc+top25perc+s(f_undergrad)+s(p_undergrad)+room_board+s(book
s)+personal+ph_d+terminal+s_f_ratio+perc_alumni+expend+grad_rate, data = colleges)

#Bivariate functions (te) added in model below
gam.m3 <- gam(outstate~
s(apps)+accept+enroll+top10perc+top25perc+te(f_undergrad,p_undergrad)+room_board+s(books)
+personal+ph_d+terminal+s_f_ratio+perc_alumni+expend+grad_rate, data = colleges)

anova(gam.m1, gam.m2, gam.m3, test = "F")
## Analysis of Deviance Table
##
## Model 1: outstate ~ apps + accept + enroll + top10perc + top25perc + f_undergrad +
##     p_undergrad + room_board + books + personal + ph_d + terminal +
##     s_f_ratio + perc_alumni + expend + grad_rate
## Model 2: outstate ~ s(apps) + accept + enroll + top10perc + top25perc +
##     s(f_undergrad) + s(p_undergrad) + room_board + s(books) +
##     personal + ph_d + terminal + s_f_ratio + perc_alumni + expend +
##     grad_rate
## Model 3: outstate ~ s(apps) + accept + enroll + top10perc + top25perc +
##     te(f_undergrad, p_undergrad) + room_board + s(books) + personal +
##     ph_d + terminal + s_f_ratio + perc_alumni + expend + grad_rate
##   Resid. Df Resid. Dev     Df  Deviance      F    Pr(>F)
## 1    547.00 2092185295
## 2    537.48 1930186071 9.5164 161999224 4.8130 1.988e-06 ***
## 3    531.50 1888350607 5.9815  41835464 1.9775   0.06732 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(gam.m2)
```

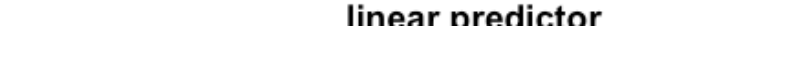


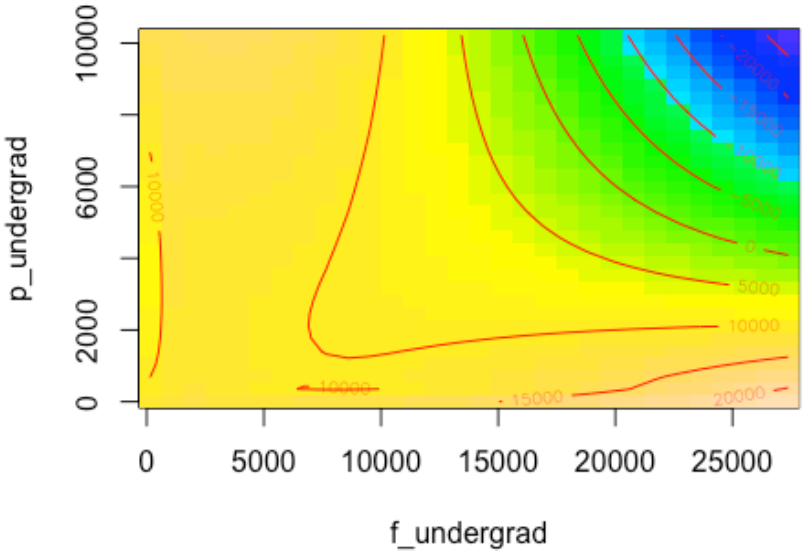







```
vis.gam(gam.m3, view = c("f_undergrad","p_undergrad"),
    plot.type = "contour", color = "topo")
```



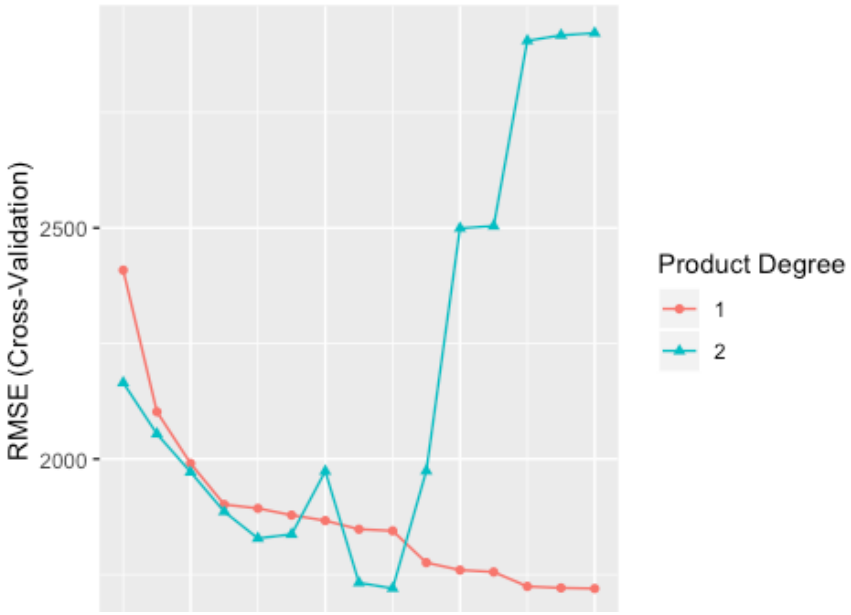


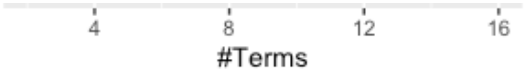
In the first model (gam.m1), all terms are treated as linear terms. S functions were incorporated into the second GAM model (gam.m2) for some predictors that exhibited possible non-linear relationships with the outcome variable (out-of-state tuition) in scatterplots in Question 1A. Lastly, a bivariate function was added to the 3rd model (gam.m3), along with s functions. The ANOVA model allows us to see if the simple, linear model (gam.m1) is sufficient to explain the association between the predictors and out of state tuition. The second model has a p-value of 1.99x10^6 and the third model has a p-value of 0.67. The second model is significant, and thus we choose this model over the simple linear model.

QUESTION 1D

Fit a multivariate adaptive regression spline (MARS) model using all the predictors. Report the final model. Present the partial dependence plot of an arbitrary predictor in your finalmodel.

```
ctrl1 <- trainControl(method = "cv", number = 10)
mars_grid <- expand.grid(degree = 1:2,
                         nprune = 2:16)

set.seed(2)
mars.fit <- train(x, y,
                  method = "earth",
                  tuneGrid = mars_grid,
                  trControl = ctrl1)

ggplot(mars.fit)
```





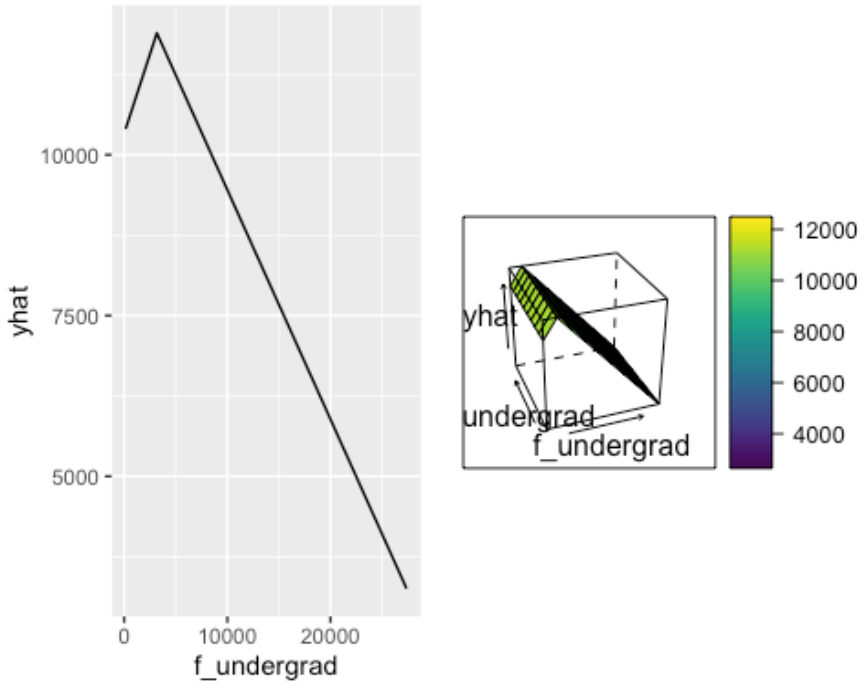

```
mars.fit$bestTune
##    nprune degree
## 15     16      1
coef(mars.fit$finalModel)
##         (Intercept)     h(expend-15365)  h(4450-room_board)
##       11157.3323100          -0.6964270          -1.2721516
##     h(grad_rate-97)     h(97-grad_rate) h(f_undergrad-1355)
##        -242.9852028         -24.1380627          -0.3567241
## h(1355-f_undergrad)   h(22-perc_alumni)        h(apps-3712)
##          -1.7564383         -77.0359905           7.0815293
##    h(1300-personal)      h(913-enroll)     h(2193-accept)
##           1.0492662           5.2944664          -1.9951560
##     h(expend-6881)       h(apps-3877)   h(s_f_ratio-10.1)
##           0.6896165          -6.7393719         -97.7224105
##   h(s_f_ratio-17.8)
##         222.5913868
```

The optimal model includes 16 terms (includes intercept) and 1 degree of freedom. The final model is reported below:

(Intercept) 11157.3323100
h(expend-15365) -0.6964270
h(4450-room_board) -1.2721516
h(grad_rate-97) -242.9852028 h(97-grad_rate) -24.1380627
h(f_undergrad-1355) -0.3567241 h(1355-f_undergrad) -1.7564383
h(22-perc_alumni) -77.0359905 h(apps-3712) 7.0815293 h(1300-personal) 1.0492662
h(913-enroll) 5.2944664 h(2193-accept) -1.9951560 h(expend-6881) 0.6896165
h(apps-3877) -6.7393719
h(s_f_ratio-10.1) -97.7224105 h(s_f_ratio-17.8) 222.5913868

To better understand the relationship between these features and out of state tuition, we can create partial dependence plots (PDPs) for each feature individually (f_undergrad is used below) and also an interaction PDP (f_undergrad and p_undergrad). This is used to examine the marginal effects of predictors.

```
#partial dependence plot of an abritrary predictor:
p1 <- partial(mars.fit, pred.var = c("f_undergrad"), grid.resolution = 10) %>% autoplot()

#Interaction partial dependence plot:
p2 <- partial(mars.fit, pred.var = c("f_undergrad", "p_undergrad"), grid.resolution = 10) %>%
    plotPartial(levelplot = FALSE, zlab = "yhat", drape = TRUE,
                screen = list(z = 20, x = -60))

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



QUESTION 1E

Based on the above GAM and MARS models, predict the out-of-state tuition of Columbia University.

```
test_columbia=
 read_excel(path = "./data/college.xlsx", sheet = 1) %>%
 janitor::clean_names() %>%
 filter(college == "Columbia University") %>%
 select(college, apps, accept, enroll, top10perc, top25perc, f_undergrad, p_undergrad, room_board,
books, personal, ph_d, terminal, s_f_ratio, perc_alumni, expend, grad_rate, outstate)

pred1 <- predict(gam.m2, newdata = test_columbia)
pred1
##        1
## 19149.33
pred2 <- predict(mars.fit, newdata= test_columbia)
pred2
##            y
## [1,] 18520.5
```

Based on the second generalized additive model in Questions 1C (gam.m2), the out-of-state tuition for Columbia University is predicted to be $19,149.33.

Based on the MARS model in Questions 1D, the out-of-state tuition for Columbia University is predicted to be $18,520.5.