ds2_hw2

Ruilian Zhang

3/6/2022

Contents


```
# data cleaning
df = read_csv("College.csv") %>%
  janitor::clean_names() %>%
  select(-college) %>%
  select(outstate, everything()) %>%
  na.omit()
```

```
## Rows: 565 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): College
## dbl (17): Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergr...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# data partition
indexTrain = createDataPartition(y = df$outstate,
                                 p = 0.8,
                                 list = FALSE)


train_df = df[indexTrain, ]
test_df = df[-indexTrain, ]

x_train = model.matrix(outstate ~ ., train_df)[ , -1]
y_train = train_df$outstate

x_test = model.matrix(outstate ~ ., test_df)[ , -1]
y_test = test_df$outstate
```

Exploratory data analysis (using train_df)

```
# data dimension and summary
dim(train_df)
```

```
## [1] 453  17
```

```
summary(train_df)
```

```
##     outstate          apps           accept          enroll
##  Min.   : 2340   Min.   :  141   Min.   :  118   Min.   :  46.0
##  1st Qu.: 9100   1st Qu.:  632   1st Qu.:  501   1st Qu.: 211.0
##  Median :11200   Median : 1151   Median :  872   Median : 334.0
##  Mean   :11783   Mean   : 2036   Mean   : 1346   Mean   : 474.1
##  3rd Qu.:13960   3rd Qu.: 2220   3rd Qu.: 1610   3rd Qu.: 527.0
##  Max.   :21700   Max.   :20192   Max.   :13007   Max.   :4615.0
##    top10perc       top25perc      f_undergrad     p_undergrad
##  Min.   : 1.00   Min.   :  9.00   Min.   :  199   Min.   :    1.0
##  1st Qu.:17.00   1st Qu.: 42.00   1st Qu.:  847   1st Qu.:   63.0
##  Median :25.00   Median : 56.00   Median : 1306   Median :  211.0
##  Mean   :29.47   Mean   : 57.24   Mean   : 1929   Mean   :  451.7
##  3rd Qu.:36.00   3rd Qu.: 70.00   3rd Qu.: 2128   3rd Qu.:  542.0
##  Max.   :96.00   Max.   :100.00   Max.   :27378   Max.   :10221.0
##    room_board       books           personal         ph_d
##  Min.   :2370   Min.   : 250.0   Min.   : 300   Min.   :  8.00
##  1st Qu.:3750   1st Qu.: 450.0   1st Qu.: 800   1st Qu.: 59.00
##  Median :4400   Median : 500.0   Median :1100   Median : 74.00
##  Mean   :4601   Mean   : 552.1   Mean   :1211   Mean   : 71.25
##  3rd Qu.:5420   3rd Qu.: 600.0   3rd Qu.:1500   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :100.00
##     terminal        s_f_ratio      perc_alumni        expend
##  Min.   : 24.00   Min.   : 2.90   Min.   : 2.00   Min.   : 3480
##  1st Qu.: 68.00   1st Qu.:11.20   1st Qu.:16.00   1st Qu.: 7444
##  Median : 81.00   Median :12.80   Median :25.00   Median : 8954
##  Mean   : 78.79   Mean   :12.98   Mean   :25.62   Mean   :10514
##  3rd Qu.: 92.00   3rd Qu.:14.50   3rd Qu.:34.00   3rd Qu.:11561
##  Max.   :100.00   Max.   :39.80   Max.   :64.00   Max.   :56233
##    grad_rate
##  Min.   : 15.00
##  1st Qu.: 59.00
##  Median : 70.00
##  Mean   : 69.29
##  3rd Qu.: 81.00
##  Max.   :118.00
```

```
skimr::skim(train_df)
```

Table 1: Data summary

Name	train_df

Table 1: Data summary

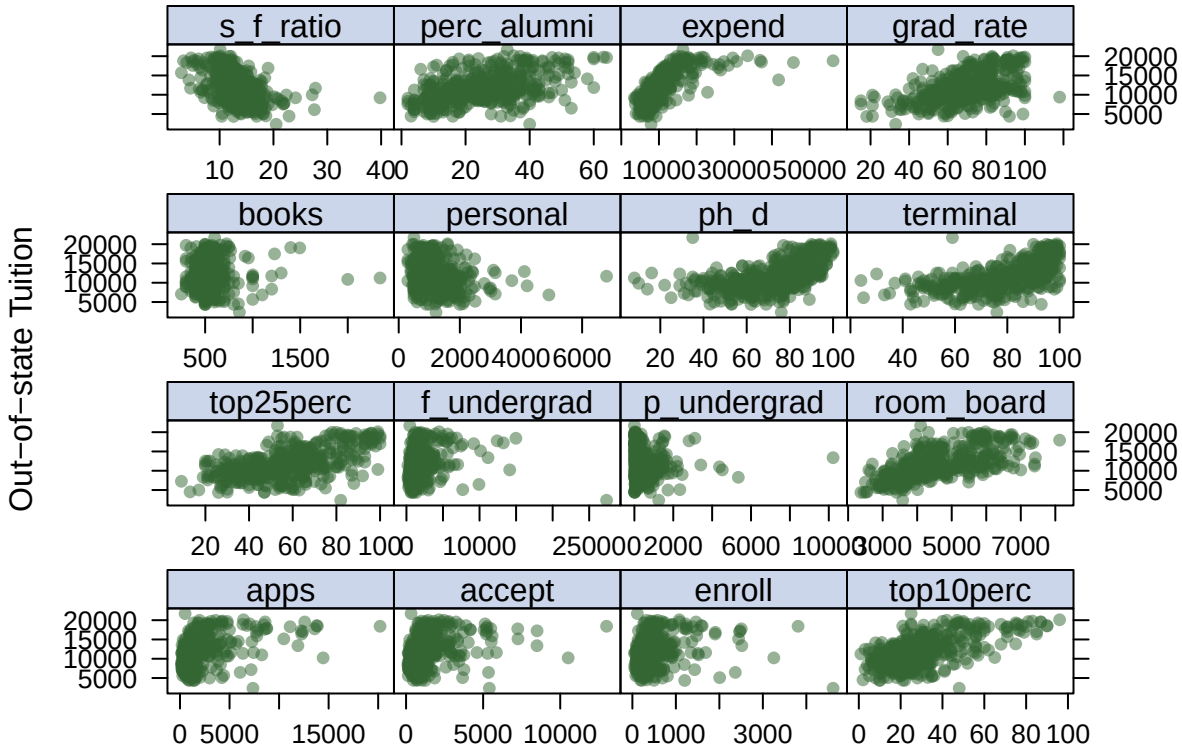
Number of rows	453
Number of columns	17
Column type frequency:	
numeric	17
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
outstate	0	1	11783.02	3747.46	2340.0	9100.0	11200.0	13960.0	21700.0	
apps	0	1	2035.86	2530.96	141.0	632.0	1151.0	2220.0	20192.0	
accept	0	1	1346.27	1443.06	118.0	501.0	872.0	1610.0	13007.0	
enroll	0	1	474.14	489.03	46.0	211.0	334.0	527.0	4615.0	
top10perc	0	1	29.47	17.66	1.0	17.0	25.0	36.0	96.0	
top25perc	0	1	57.24	19.78	9.0	42.0	56.0	70.0	100.0	
f_undergrad	0	1	1929.45	2242.99	199.0	847.0	1306.0	2128.0	27378.0	
p_undergrad	0	1	451.67	770.80	1.0	63.0	211.0	542.0	10221.0	
room_board	0	1	4601.44	1097.71	2370.0	3750.0	4400.0	5420.0	8124.0	
books	0	1	552.06	185.44	250.0	450.0	500.0	600.0	2340.0	
personal	0	1	1210.94	647.72	300.0	800.0	1100.0	1500.0	6800.0	
ph_d	0	1	71.25	17.52	8.0	59.0	74.0	85.0	100.0	
terminal	0	1	78.79	15.78	24.0	68.0	81.0	92.0	100.0	
s_f_ratio	0	1	12.98	3.57	2.9	11.2	12.8	14.5	39.8	
perc_alumni	0	1	25.62	12.55	2.0	16.0	25.0	34.0	64.0	
expend	0	1	10513.60	5670.20	3480.0	7444.0	8954.0	11561.0	56233.0	
grad_rate	0	1	69.29	16.72	15.0	59.0	70.0	81.0	118.0	

There are 453 rows and 17 columns in training data, all the variables are numeric.

```
# set plot theme
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)

# scatter plot
# all predictors are included since they are all continuous
featurePlot(
  x_train,
  y_train,
  plot = "scatter",
  labels = c("","Out-of-state Tuition"),
  layout = c(4, 4))
```



From the scatter plot above, we can see that there might be some linear trends between the outcome variable `outstate` and some of the predictors, for example, `phd` and `terminal`.

Smoothing splines

```
set.seed(2570)

# fit smoothing spline models using terminal as the only predictor of outstate
fit_ss = smooth.spline(x = train_df$terminal, y = train_df$outstate)

# optimal degree of freedom obtained by generalized cross-validation
fit_ss$df
```

```
## [1] 4.382707
```

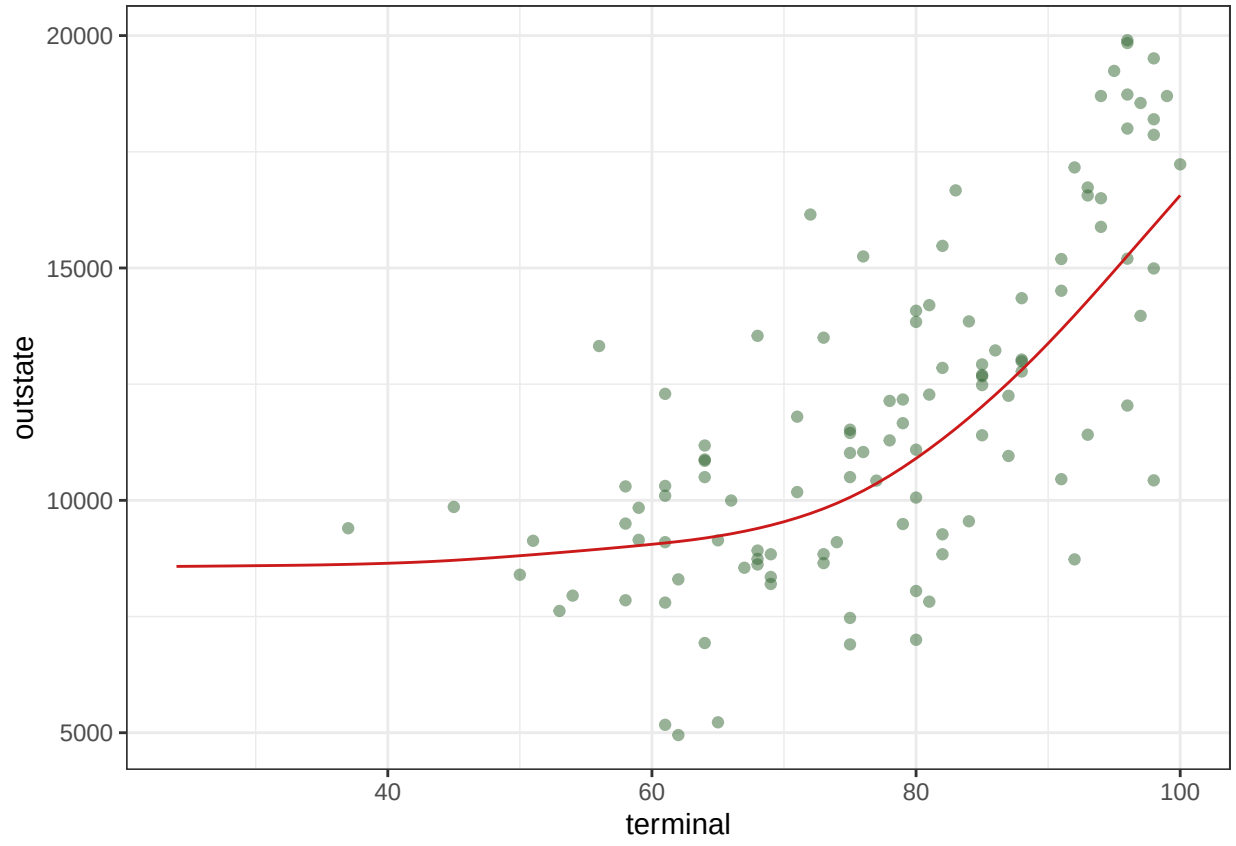
The optimal degree of freedom obtained by default cross validation is 4.383.

Use this **optimal degree of freedom** to make following predictions:

```
# make prediction using a grid of terminal values
# generate predictor grid
range(train_df$terminal)
```

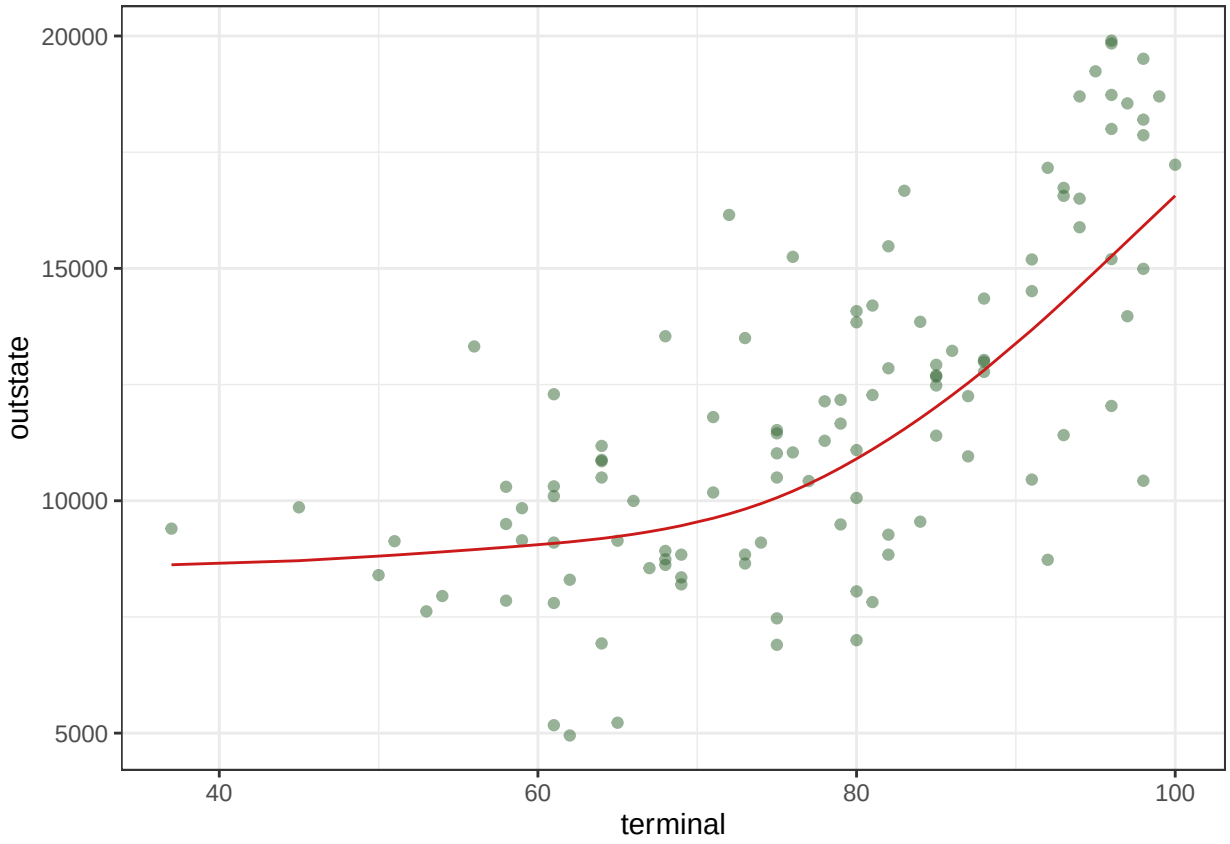
```
## [1]  24 100
```

```
terminal_grid <- seq(from = 24, to = 100, by = 1)

# make prediction
pred_ss = predict(fit_ss,
                  x = terminal_grid)

pred_ss_df = data.frame(predicted = pred_ss$y,
                        terminal = terminal_grid)

# plot test data
p = ggplot(data = test_df, aes(x = terminal, y = outstate)) +
     geom_point(color = rgb(.2, .4, .2, .5))

# plot predicted value
p +
  geom_line(aes(x = terminal, y = predicted),
            data = pred_ss_df,
            color = rgb(.8, .1, .1, 1)) +
  theme_bw()
```




```
# make prediction using test data
pred_ss_test = predict(fit_ss,
                       x = test_df$terminal)

pred_ss_test_df = data.frame(predicted = pred_ss_test$y,
```

```
                              terminal = test_df$terminal)

# plot predicted value
p +
  geom_line(aes(x = terminal, y = predicted),
            data = pred_ss_test_df,
            color = rgb(.8, .1, .1, 1)) +
  theme_bw()
```

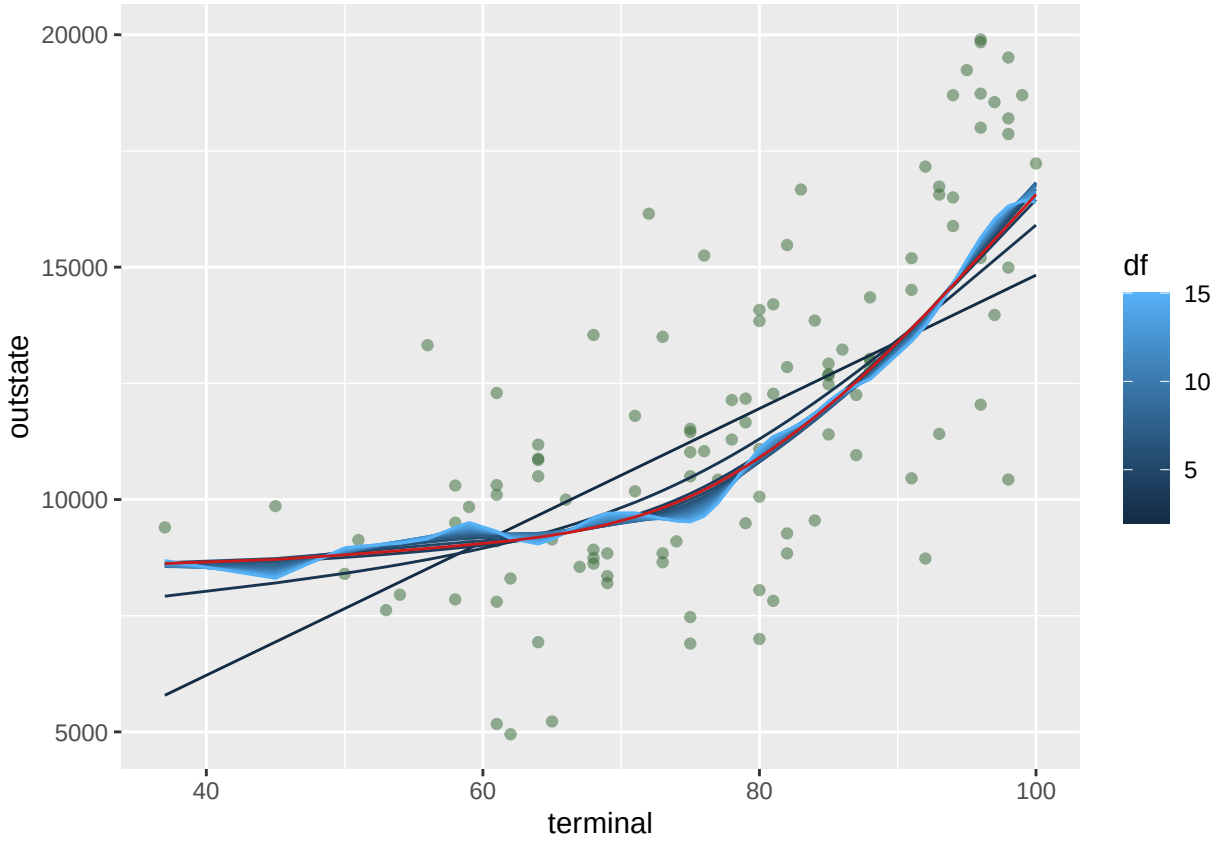


Use **a range of degree of freedom** to make predictions:

```
# write a function using a range of df to fit models
ss_func = function(df) {

  fit_ss_fun = smooth.spline(x = train_df$terminal,
                                    y = train_df$outstate,
                                    df = df)

  pred_ss_fun = predict(fit_ss_fun, x = test_df$terminal)


  pred_ss_df_fun = data.frame(predicted = pred_ss_fun$y,
                           terminal = test_df$terminal,
                           df = df)

}
```

```r
# create a list of df
# 1 < df <= 16 - 1
df_list = seq(2, 15, 1)

# run the function using df_list
output_ss = list()

for (x in df_list) {
  output_ss[[x]] = ss_func(x)
}

# do.call() executes a function by its name and a list of corresponding arguments
# e.g. do.call("any_function", arguments_list)
output_ss_df = do.call("rbind", output_ss) %>%
  as.data.frame()
```

```r
# plot results for a range of df
p +
  geom_line(aes(x = terminal, y = predicted, group = df, color = df), data = output_ss_df) +
  geom_line(aes(x = terminal, y = predicted), data = pred_ss_test_df, color = rgb(.8, .1, .1, 1))
```



The above plot shows the fitted smoothing spline models using a range of degree of freedoms. The lines wiggle around the red line, which is the model using the optimum degree of freedom.
As the degree of freedom approaching to 2, the line gets more linear; as the degree of freedom approaching to 15, the line gets more wiggled.

Among all the fitted lines within the (2, 15) degree of freedom range, df = 4.383 should be the nearest to the red line.

Generalized Additive Model (GAM)

```
set.seed(2570)

# set cross validation method
ctrl = trainControl(method = "cv", number = 10)

# fit a GAM model using all the predictors
# ngcv package not available for current R version, siwth to caret
gam_fit = train(x = x_train,
                y = y_train,
                method = "gam",
                #tuneGrid = data.frame(method = "GCV.Cp",select = c(TRUE, FALSE)),
                trControl = ctrl)
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-39. For overview type 'help("mgcv-package")'.
```

```
gam_fit$bestTune
```

```
##   select method
## 2   TRUE GCV.Cp
```

```
gam_fit$finalModel
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(perc_alumni) + s(terminal) + s(books) + s(top10perc) +
##     s(ph_d) + 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)
##
## Estimated degrees of freedom:
```

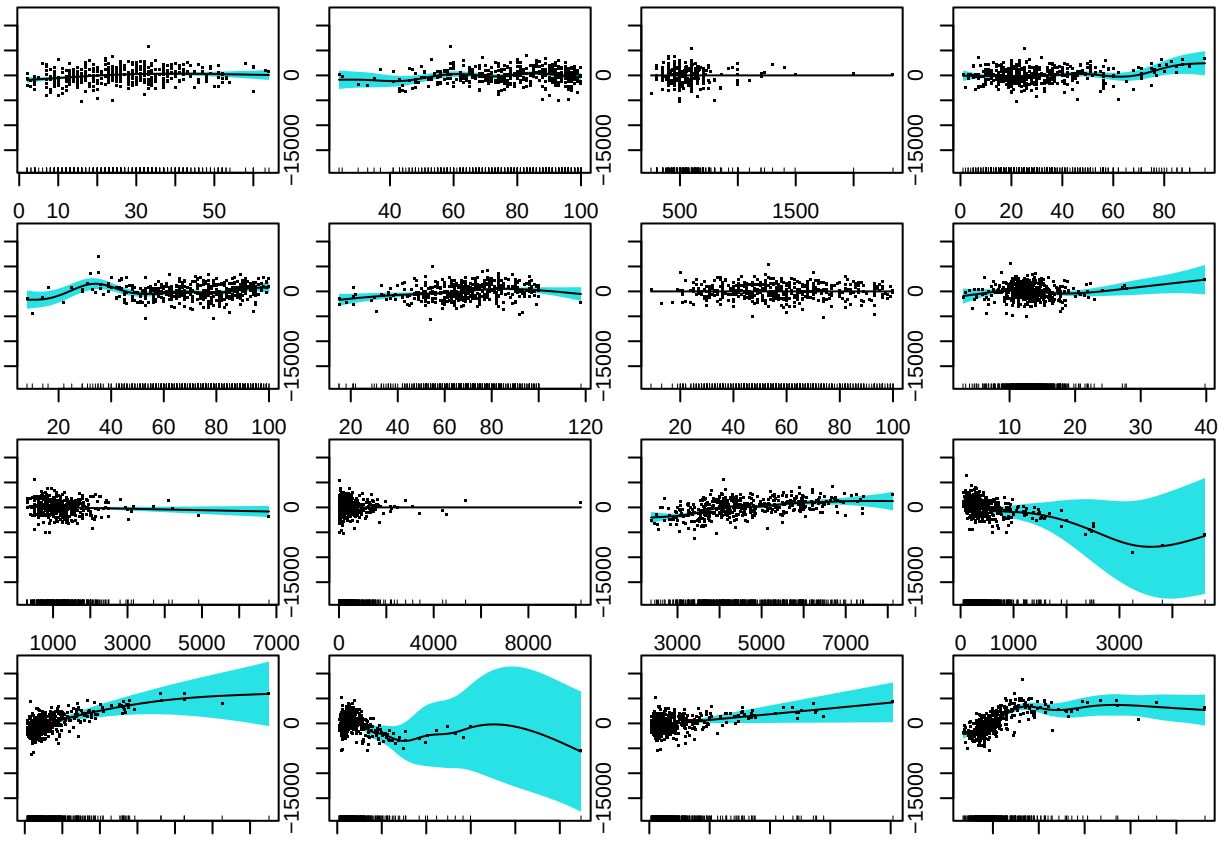
```
## 2.106 5.820 0.000 5.795 6.883 3.512 0.000
## 3.793 0.627 0.000 5.663 3.613 1.857 6.840
## 0.696 6.206  total = 54.41
##
## GCV score: 2648846
```

```
summary(gam_fit)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(perc_alumni) + s(terminal) + s(books) + s(top10perc) +
##     s(ph_d) + 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) 11783.02      71.73   164.3   <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) 2.106e+00      9  1.768 0.000116 ***
## s(terminal)    5.820e+00      9  1.448 0.023822 *
## s(books)       1.055e-05      9  0.000 0.418641
## s(top10perc)   5.795e+00      9  1.027 0.112380
## s(ph_d)        6.883e+00      9  2.914 0.000187 ***
## s(grad_rate)   3.512e+00      9  2.609 1.62e-05 ***
## s(top25perc)   5.065e-07      9  0.000 0.831478
## s(s_f_ratio)   3.793e+00      9  1.254 0.009830 **
## s(personal)    6.270e-01      9  0.246 0.057396 .
## s(p_undergrad) 2.768e-06      9  0.000 0.755315
## s(room_board)  5.663e+00      9  8.215  < 2e-16 ***
## s(enroll)      3.613e+00      9  0.896 0.017352 *
## s(accept)      1.857e+00      9  1.903 5.53e-06 ***
## s(f_undergrad) 6.840e+00      9  3.167 3.55e-06 ***
## s(apps)        6.958e-01      9  0.495 0.006274 **
## s(expend)      6.206e+00      9 15.438  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.834   Deviance explained = 85.4%
## GCV = 2.6488e+06  Scale est. = 2.3307e+06  n = 453
```

```
# plot the results
par(mar=c(1,1,1,1))
par(mfrow = c(4, 4))

plot(gam_fit$finalModel,
```

```
    residuals = TRUE,
    all.terms = TRUE,
    shade = TRUE,
    shade.col = 5)
```




```
# train RMSE of final model
gam_train_rmse = sqrt(mean((y_train - predict(gam_fit)) ^ 2))
gam_train_rmse
```

```
## [1] 1432.036
```

```
# make predictions
gam_pred = predict(gam_fit, x_test)

# test RMSE of final model
gam_test_rmse = sqrt(mean(y_test - gam_pred) ^ 2)
gam_test_rmse
```

```
## [1] 122.8602
```

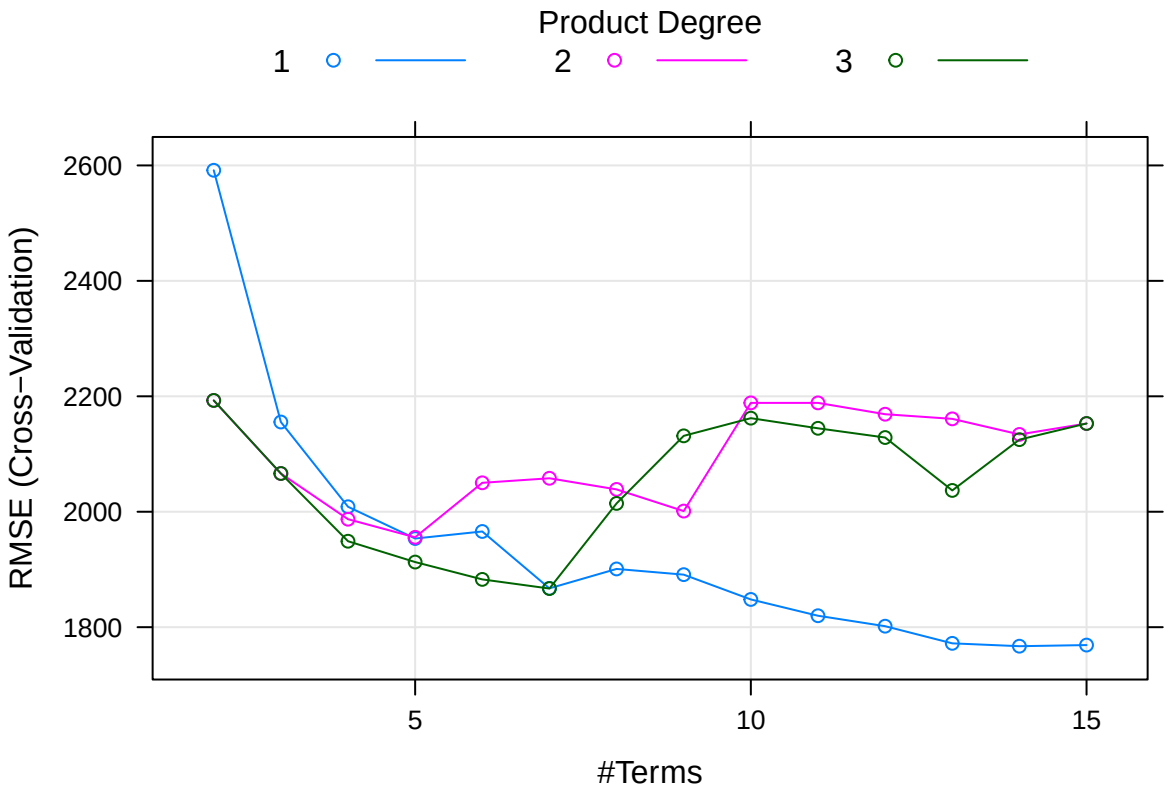
The training RMSE is1432.0359678 and the test RMSE is 122.8602138.
Coefficients are not printed for smooth terms because each smooth term has several coefficients corresponding to different basis functions. The degrees of freedom of each term represent the complexity of the smooth function.
In the final model, `perc_alumni`, `grad_rate`, `room_board`, `enroll`, `accept`, `f_undergrad`, and `expend` are the most significant terms.

Multivariate Adaptive Regression spline (MARS)

```
set.seed(2570)

# generate all possible combinations of degree and prune
mars_grid = expand.grid(degree = 1:3,
                        nprune = 2:15)

# fit MARS model using all predictors
mars_fit = train(x = x_train,
                 y = y_train,
                 method = "earth",
                 tuneGrid = mars_grid,
                 trControl = ctrl)
```

```
## Loading required package: earth

## Loading required package: Formula

## Loading required package: plotmo

## Loading required package: plotrix

## Loading required package: TeachingDemos
```

```
# plot results
plot(mars_fit)
```




```
mars_fit$bestTune
```

```
##    nprune degree
## 13     14      1
```

```
summary(mars_fit$finalModel)
```

```
## Call: earth(x=matrix[453,16], y=c(7440,12280,11...), keepxy=TRUE, degree=1,
##             nprune=14)
##
##                      coefficients
## (Intercept)           14849.6502
## h(apps-1891)              0.4424
## h(1580-accept)           -1.5022
## h(876-enroll)             3.9629
## h(1419-f_undergrad)      -1.6289
## h(f_undergrad-1419)      -0.3937
## h(4150-room_board)       -1.4941
## h(room_board-4150)        0.4274
## h(ph_d-82)               79.8929
## h(10.1-s_f_ratio)      -256.8227
## h(30-perc_alumni)       -54.5362
## h(15954-expend)          -0.5956
## h(grad_rate-64)          76.6878
## h(grad_rate-80)        -115.3050
```

```
##
## Selected 14 of 22 terms, and 10 of 16 predictors (nprune=14)
## Termination condition: RSq changed by less than 0.001 at 22 terms
## Importance: expend, room_board, perc_alumni, accept, f_undergrad, apps, ...
## Number of terms at each degree of interaction: 1 13 (additive model)
## GCV 2773199    RSS 1110969178    GRSq 0.8029628    RSq 0.8249788
```

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

```
##         (Intercept)     h(15954-expend)     h(grad_rate-80)  h(room_board-4150)
##       14849.6501833          -0.5956494        -115.3050222           0.4273863
##  h(4150-room_board) h(f_undergrad-1419) h(1419-f_undergrad)        h(apps-1891)
##          -1.4940616          -0.3936602          -1.6289284           0.4424265
##   h(30-perc_alumni)       h(876-enroll)          h(ph_d-82)      h(1580-accept)
##         -54.5362311           3.9629029          79.8929331          -1.5021748
##     h(grad_rate-64)   h(10.1-s_f_ratio)
##          76.6878325        -256.8227058
```

```
# train RMSE of final model
mars_train_rmse = sqrt(mean((y_train - predict(mars_fit)) ^ 2))
mars_train_rmse
```

```
## [1] 1566.037
```

```
# make predictions
mars_pred = predict(mars_fit, x_test)

# test RMSE of final model
mars_test_rmse = sqrt(mean(y_test - gam_pred) ^ 2)
mars_test_rmse
```

```
## [1] 122.8602
```

The training RMSE is1566.0365868 and the test RMSE is 122.8602138.
The final model's maximum degree of interactions is 1, which means the final model is an additive model.
`nprune` is 13, which means there are 13 terms in the final model, including intercept.
The most important terms in the final model are `expend`, `room_board`, `perc_alumni`, `accept`, and `enroll`.

```
# partial dependence plot
# use `books` as predictor
pdp = pdp::partial(mars_fit,
                   pred.var = c("enroll"),
                   grid.resolution = 20) %>%
  autoplot()

# use `books` and `expend` as predictors

pdp_2d = pdp::partial(mars_fit,
                      pred.var = c("enroll", "expend"),
                      grid.resolution = 20) %>%
  pdp::plotPartial(levelplot = FALSE,
```

```
                  zlab = "yhat",
                  drape = TRUE,
                  screen = list(z = 20, x = -60))

# grid.arrange(pdp, pdp_2d, n_col = 2)
```

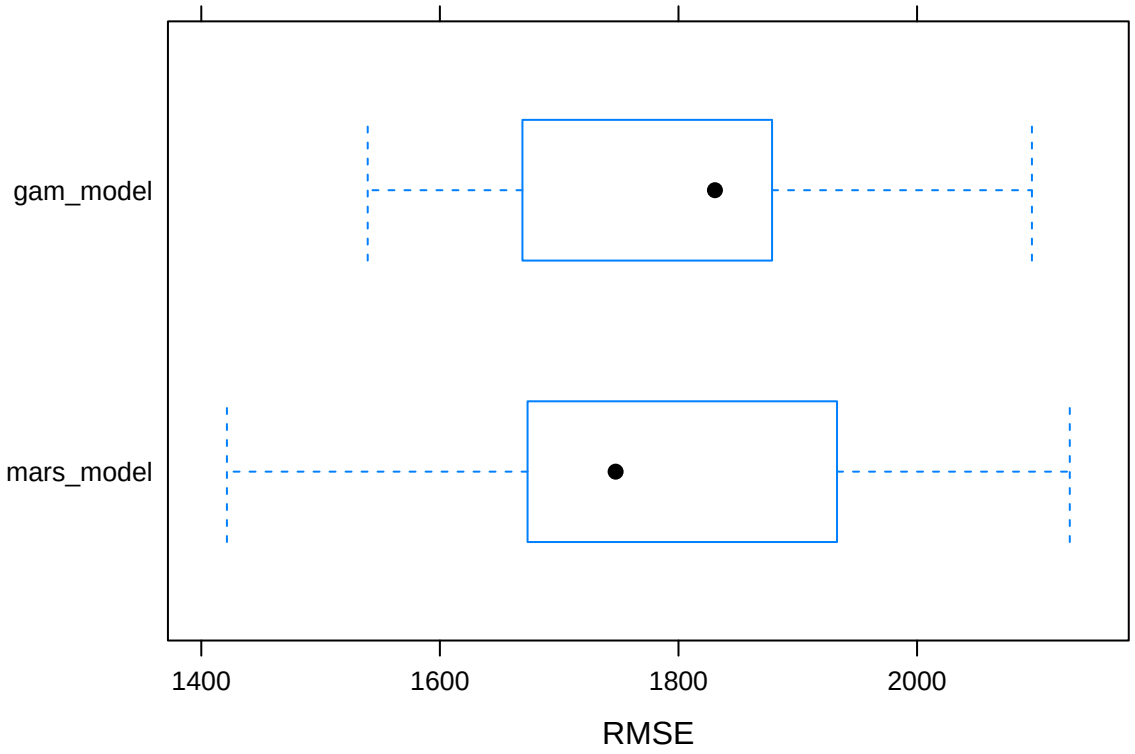
Within the range of approximately less than 1000, we can see a trend of decrease on the response variable when as `enroll` increases, and the value of response variable stays stable afterwards. This turning point is the knot.

Model selection

```
resamp = resamples(list(gam_model = gam_fit,
                        mars_model = mars_fit))

summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: gam_model, mars_model
## Number of resamples: 10
##
## MAE
##                Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## gam_model  1192.241 1309.794 1390.926 1373.967 1440.495 1543.758    0
## mars_model 1167.253 1306.795 1337.461 1385.177 1524.128 1576.808    0
##
## RMSE
##                Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## gam_model  1539.469 1702.204 1830.554 1816.208 1875.312 2096.275    0
## mars_model 1421.481 1679.524 1747.304 1767.062 1887.125 2127.992    0
##
## Rsquared
##                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## gam_model  0.6935877 0.7345419 0.7765513 0.7697549 0.8029582 0.8327107    0
## mars_model 0.6635357 0.7432341 0.7945434 0.7774462 0.8161699 0.8553172    0
```

```
bwplot(resamp, metric = "RMSE")
```



In this data example, we might prefer the use of MARS model over linear model when predicting the out-of-state tuition, since the RMSE of MARS model is smaller, which indicates the MARS model fits the data better.