HW7

Trey Davidson

2024-10-22

#### Question 6

## A.)

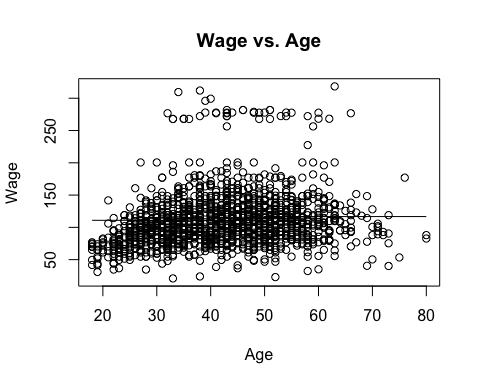
library(ISLR2)  
  
  
set.seed(10)   
train\_index <- sample(length(Wage$year), length(Wage$year) / 2)  
  
train\_data <- Wage[train\_index, ]  
test\_data <- Wage[-train\_index, ]  
  
  
RSS <- numeric(10)  
for (d in 1:10)  
   
{  
   
   
 model <- lm(wage ~ poly(age, degree = d), data = train\_data)  
   
   
 predictions <- predict(model, newdata = test\_data)  
   
   
 RSS[d] <- sum(( test\_data$wage - predictions) ^ 2)  
  
}  
  
  
  
model <- lm(wage ~ poly(age, degree = which.min(RSS)), data = train\_data)  
  
summary(model)

##   
## Call:  
## lm(formula = wage ~ poly(age, degree = which.min(RSS)), data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -100.966 -24.190 -4.152 15.992 198.920   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 112.547 1.029 109.324 < 2e-16 \*\*\*  
## poly(age, degree = which.min(RSS))1 302.897 39.872 7.597 5.33e-14 \*\*\*  
## poly(age, degree = which.min(RSS))2 -339.751 39.872 -8.521 < 2e-16 \*\*\*  
## poly(age, degree = which.min(RSS))3 102.061 39.872 2.560 0.0106 \*   
## poly(age, degree = which.min(RSS))4 -68.190 39.872 -1.710 0.0874 .   
## poly(age, degree = which.min(RSS))5 -16.222 39.872 -0.407 0.6842   
## poly(age, degree = which.min(RSS))6 32.634 39.872 0.818 0.4132   
## poly(age, degree = which.min(RSS))7 32.171 39.872 0.807 0.4199   
## poly(age, degree = which.min(RSS))8 -21.929 39.872 -0.550 0.5824   
## poly(age, degree = which.min(RSS))9 -66.989 39.872 -1.680 0.0931 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 39.87 on 1490 degrees of freedom  
## Multiple R-squared: 0.08836, Adjusted R-squared: 0.08285   
## F-statistic: 16.05 on 9 and 1490 DF, p-value: < 2.2e-16

THis model is overal statsticl sifinfict but certain degrees are not ioke 5th 6th 7th and 8th.

## B.)

library(ISLR2)  
  
  
set.seed(10)   
train\_index <- sample(length(Wage$year), length(Wage$year) / 2)  
  
train\_data <- Wage[train\_index, ]  
test\_data <- Wage[-train\_index, ]  
  
RSS <- numeric(10)  
for (num\_cuts in 1:10)  
{  
 age\_cut <- cut(Wage$age, breaks = num\_cuts + 1)  
   
 model <- glm(wage ~ age\_cut, data = Wage)  
   
 predictions <- suppressWarnings(predict(model, newdata = test\_data))  
   
 RSS[num\_cuts] <- sum((test\_data$wage - predictions) ^ 2)  
}  
  
  
  
optimal\_cuts <- which.min(RSS)  
  
  
model <- glm(wage ~ cut(age, breaks = optimal\_cuts + 1), data = train\_data)  
  
  
test\_data$age\_cut <- cut(test\_data$age, breaks = optimal\_cuts + 1, include.lowest = TRUE)  
predictions <- predict(model, newdata = test\_data)  
  
  
plot(test\_data$age, test\_data$wage,   
 xlab = "Age",   
 ylab = "Wage",   
 main = "Wage vs. Age")  
  
  
lines(test\_data$age[order(test\_data$age)],   
 predictions[order(test\_data$age)])



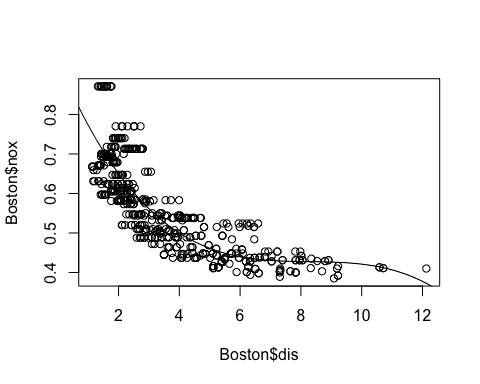
The optimal cuts lokked to be 2 ### Question 9

## A.)

library(ISLR2)  
  
  
poly\_model <- lm(nox ~ poly(dis, 3), data = Boston)  
  
  
plot(Boston$dis, Boston$nox)  
  
  
summary(poly\_model)

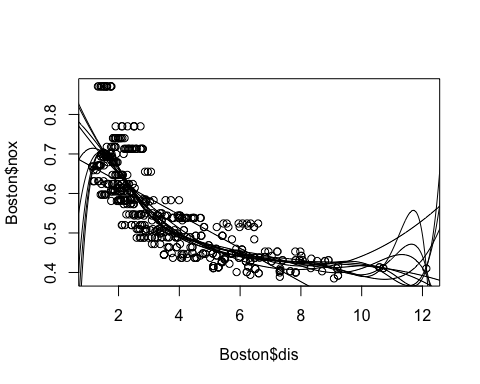
##   
## Call:  
## lm(formula = nox ~ poly(dis, 3), data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.121130 -0.040619 -0.009738 0.023385 0.194904   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.554695 0.002759 201.021 < 2e-16 \*\*\*  
## poly(dis, 3)1 -2.003096 0.062071 -32.271 < 2e-16 \*\*\*  
## poly(dis, 3)2 0.856330 0.062071 13.796 < 2e-16 \*\*\*  
## poly(dis, 3)3 -0.318049 0.062071 -5.124 4.27e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06207 on 502 degrees of freedom  
## Multiple R-squared: 0.7148, Adjusted R-squared: 0.7131   
## F-statistic: 419.3 on 3 and 502 DF, p-value: < 2.2e-16

# Creates a line that goes over data set  
dis\_seq <- seq(0, 13, length.out = 100 \* 13)  
  
predicted\_nox <- predict(poly\_model, newdata = data.frame(dis = dis\_seq))  
  
  
lines(dis\_seq, predicted\_nox)



## B.)

library(ISLR2)  
  
  
  
  
  
plot(Boston$dis, Boston$nox)  
  
  
RSS <- vector("numeric", length = 10)  
  
  
for (d in 1:10)  
{  
 poly\_model <- lm(nox ~ poly(dis, d), data = Boston)  
   
 dis\_seq <- seq(0, 13, length.out = 100 \* 13)  
   
 predicted\_nox <- predict(poly\_model, newdata = data.frame(dis = dis\_seq))  
   
 RSS[d] = sum(poly\_model$residuals^2)  
   
 lines(dis\_seq, predicted\_nox)  
   
  
}

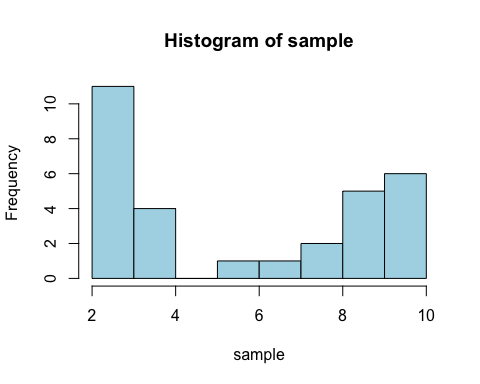


RSS

## [1] 2.768563 2.035262 1.934107 1.932981 1.915290 1.878257 1.849484 1.835630  
## [9] 1.833331 1.832171

## C.)

### Code creates 30 combinations of train and test data, and then creates a model using   
library(ISLR2)  
  
  
data <- Boston  
  
sample <- numeric(30)  
  
for (n in 1:30)  
{  
 set.seed(n)  
   
   
 train\_index <- sample(nrow(Boston), nrow(Boston) / 2, replace = FALSE)  
 train\_data <- data[train\_index, ]  
 vald\_data <- data[-train\_index, ]  
   
 RSS <- numeric(10)  
 for (d in 1:10) {  
   
 poly\_model <- lm(nox ~ poly(dis, d), data = train\_data)  
   
   
 predicted\_nox <- predict(poly\_model, newdata = vald\_data)  
   
   
 RSS[d] <- sum((vald\_data$nox - predicted\_nox)^2)  
 }  
   
 sample[n] <- which.min(RSS)  
  
}  
  
hist(sample, breaks = 10, col="lightblue")



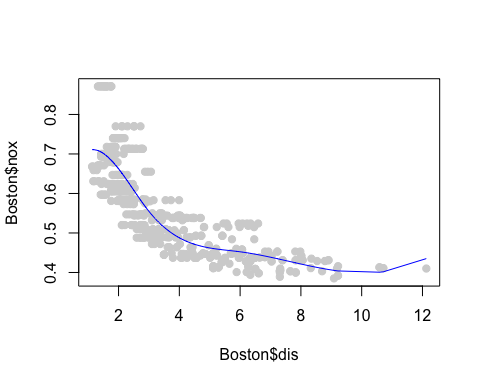
sample

## [1] 4 4 9 8 3 9 3 3 6 2 3 3 10 3 4 3 3 9 10 10 9 3 10 8 4  
## [26] 9 10 3 7 10

The histogram clearly shows that 3 degrees with being the best in 33% of validation sets. The next best 10 makes up 16%, which is a signfinct dropoff

## D.)

library(ISLR2)  
library(splines)  
  
data(Boston)  
  
  
spline <- bs(Boston$dis, degree = 3, knots = c(3, 6))  
  
  
model <- lm(Boston$nox ~ spline)  
  
  
predicted\_y <- predict(model)  
  
  
plot(Boston$dis, Boston$nox, col = "lightgrey", pch = 19)  
  
x <- sort(Boston$dis)  
y <- predicted\_y[order(Boston$dis)]  
lines(x, y, col = "blue")



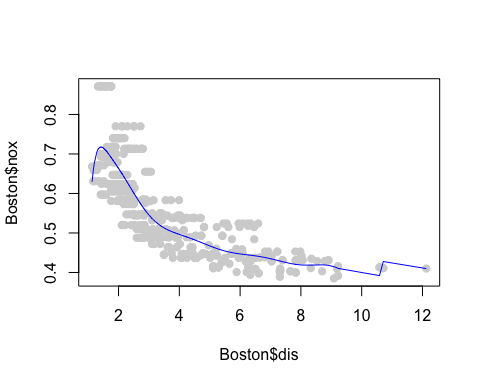
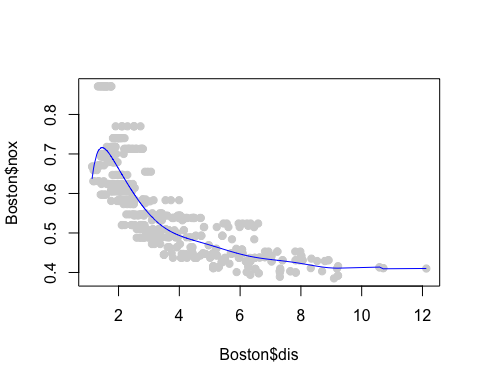
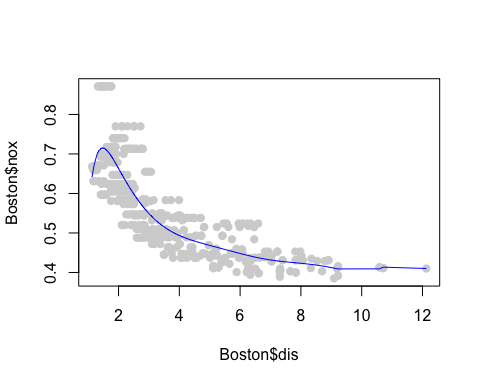
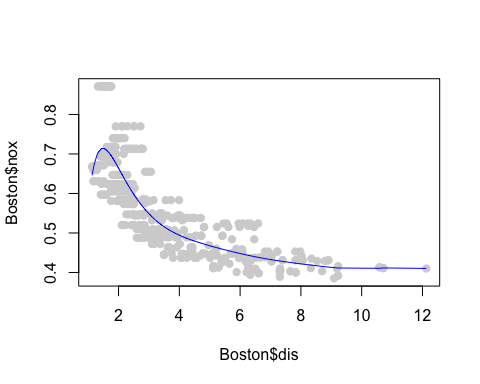
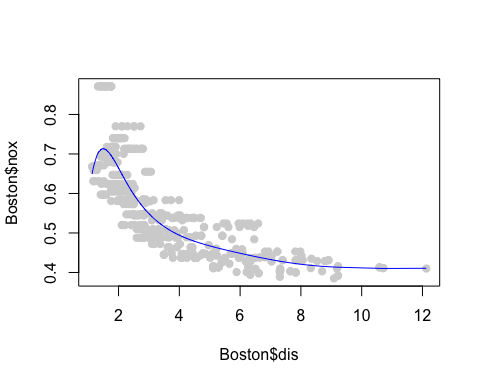
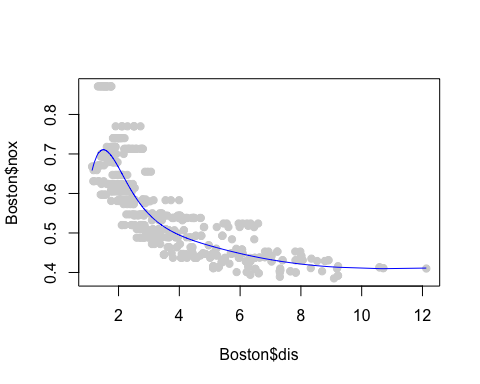
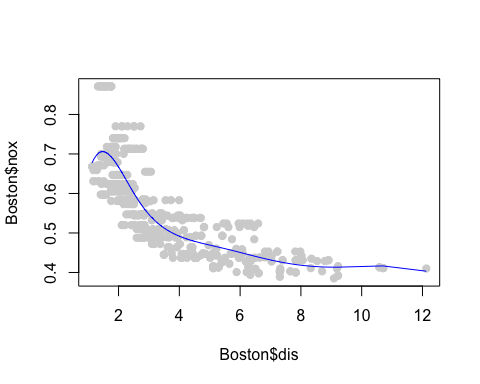
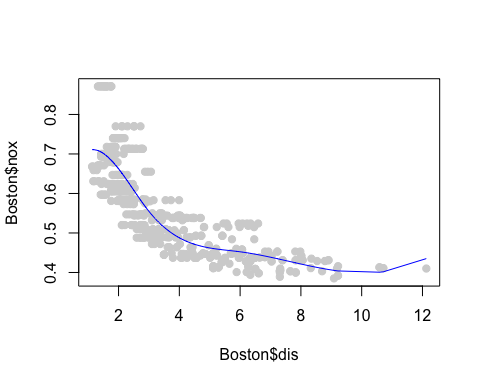
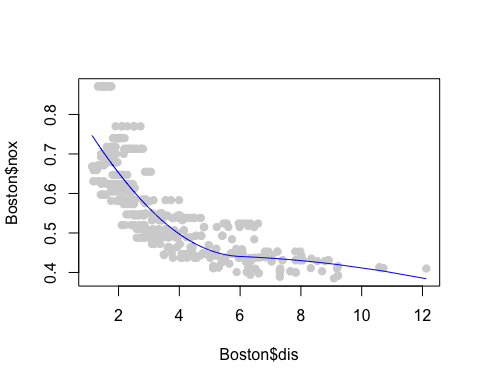
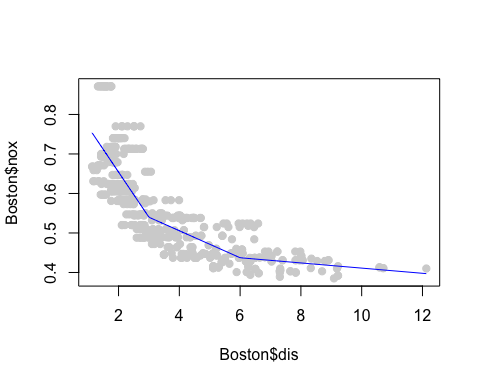
RSS <- sum(model$residuals^2)  
  
  
RSS

## [1] 1.871844

I choose the places here it looks like the line was changing. I also adjusted degree rather then df because I kept getting the same lines, and I understand that df is affected by the degree polynomial

## E.)

library(ISLR2)  
library(splines)  
  
data(Boston)  
  
  
  
  
RSS <- numeric(10)  
  
for (d in 1:10)  
   
{  
 spline <- bs(Boston$dis, degree = d, knots = c(3, 6))  
   
   
 model <- lm(Boston$nox ~ spline)  
   
   
 predicted\_y <- predict(model)  
   
   
 plot(Boston$dis, Boston$nox, col = "lightgrey", pch = 19)  
   
 x <- sort(Boston$dis)  
 y <- predicted\_y[order(Boston$dis)]  
 lines(x, y, col = "blue")  
   
   
 RSS[d] <- sum(model$residuals^2)  
   
  
}



RSS

## [1] 1.911467 1.927383 1.871844 1.841831 1.834611 1.832461 1.831092 1.830310  
## [9] 1.829565 1.825522

This results in all of them being the exact same RSS, which makes sense if the

library(ISLR2)  
library(splines)  
  
data(Boston)  
  
  
set.seed(1)  
  
  
train\_indices <- sample(1:nrow(Boston), size = 0.5 \* nrow(Boston))  
train\_data <- Boston[train\_indices, ]  
test\_data <- Boston[-train\_indices, ]  
  
RSS <- numeric(10)  
  
for (d in 1:10) {  
   
   
 spline <- bs(train\_data$dis, degree = d, knots = c(3, 6))  
   
  
 model <- lm(train\_data$nox ~ spline)  
   
  
 test\_spline <- bs(test\_data$dis, degree = d, knots = c(3, 6))  
 predicted\_y <- predict(model, newdata = data.frame(spline = test\_spline))  
   
  
 RSS[d] <- sum((predicted\_y - test\_data$nox)^2)  
   
  
   
}  
  
RSS

## [1] 5.880153 5.871868 5.856774 5.879665 5.889215 5.897870 5.896387 5.928393  
## [9] 5.912293 5.915298

It looks like the 3 degree or df 4 did the best.

### 10.

college\_data <- College  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked \_by\_ '.GlobalEnv':  
##   
## Boston

## The following object is masked from 'package:ISLR2':  
##   
## Boston

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

set.seed(10)   
train\_index <- createDataPartition(college\_data$Outstate, p = 0.8, list = FALSE)  
train\_set <- college\_data[train\_index, ]  
test\_set <- college\_data[-train\_index, ]  
  
  
initial\_model <- lm(Outstate ~ 1, data = train\_set)   
full\_model <- lm(Outstate ~ ., data = train\_set)  
  
stepwise\_model <- stepAIC(initial\_model, scope = list(lower = initial\_model, upper = full\_model), direction = "forward")

## Start: AIC=10349.19  
## Outstate ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + Room.Board 1 4634370503 5.5407e+09 9972.5  
## + Expend 1 4502160332 5.6729e+09 9987.2  
## + S.F.Ratio 1 3340801180 6.8342e+09 10103.2  
## + perc.alumni 1 3312754062 6.8623e+09 10105.8  
## + Grad.Rate 1 3137161484 7.0379e+09 10121.5  
## + Top10perc 1 3018223034 7.1568e+09 10132.0  
## + Private 1 2971416633 7.2036e+09 10136.0  
## + Top25perc 1 2332577528 7.8425e+09 10189.0  
## + Terminal 1 1810340935 8.3647e+09 10229.1  
## + PhD 1 1548410698 8.6266e+09 10248.3  
## + Personal 1 752751290 9.4223e+09 10303.3  
## + P.Undergrad 1 584245008 9.5908e+09 10314.4  
## + F.Undergrad 1 482408794 9.6926e+09 10320.9  
## + Enroll 1 260140141 9.9149e+09 10335.1  
## <none> 1.0175e+10 10349.2  
## + Apps 1 21795227 1.0153e+10 10349.9  
## + Books 1 19466133 1.0156e+10 10350.0  
## + Accept 1 14261636 1.0161e+10 10350.3  
##   
## Step: AIC=9972.52  
## Outstate ~ Room.Board  
##   
## Df Sum of Sq RSS AIC  
## + perc.alumni 1 1454187166 4086491968 9784.9  
## + Expend 1 1424670807 4116008327 9789.3  
## + S.F.Ratio 1 1212135049 4328544085 9820.7  
## + Private 1 1112778471 4427900663 9834.8  
## + Top10perc 1 962616199 4578062935 9855.6  
## + Grad.Rate 1 865368309 4675310825 9868.7  
## + Top25perc 1 678472211 4862206924 9893.1  
## + P.Undergrad 1 387513739 5153165395 9929.3  
## + F.Undergrad 1 283438968 5257240166 9941.8  
## + Terminal 1 283101079 5257578056 9941.8  
## + PhD 1 256745608 5283933526 9945.0  
## + Personal 1 239179329 5301499805 9947.0  
## + Enroll 1 167800497 5372878637 9955.4  
## + Accept 1 85928454 5454750681 9964.8  
## + Apps 1 41503613 5499175521 9969.8  
## <none> 5540679134 9972.5  
## + Books 1 13009289 5527669845 9973.1  
##   
## Step: AIC=9784.86  
## Outstate ~ Room.Board + perc.alumni  
##   
## Df Sum of Sq RSS AIC  
## + Expend 1 712475889 3374016079 9667.5  
## + S.F.Ratio 1 549637449 3536854519 9696.9  
## + Private 1 454775295 3631716673 9713.4  
## + Top10perc 1 312947858 3773544111 9737.2  
## + Grad.Rate 1 256275948 3830216020 9746.5  
## + Top25perc 1 181250441 3905241527 9758.6  
## + Terminal 1 113184459 3973307509 9769.4  
## + PhD 1 108449482 3978042487 9770.1  
## + P.Undergrad 1 97086137 3989405832 9771.9  
## + F.Undergrad 1 79650879 4006841089 9774.6  
## + Personal 1 42988949 4043503020 9780.3  
## + Enroll 1 41849942 4044642026 9780.4  
## <none> 4086491968 9784.9  
## + Accept 1 3476486 4083015483 9786.3  
## + Apps 1 1412628 4085079341 9786.6  
## + Books 1 720410 4085771558 9786.8  
##   
## Step: AIC=9667.5  
## Outstate ~ Room.Board + perc.alumni + Expend  
##   
## Df Sum of Sq RSS AIC  
## + Private 1 485953044 2888063036 9572.6  
## + Grad.Rate 1 178037377 3195978702 9635.7  
## + F.Undergrad 1 170672845 3203343235 9637.2  
## + S.F.Ratio 1 158264567 3215751512 9639.6  
## + Enroll 1 126496530 3247519549 9645.7  
## + P.Undergrad 1 119770450 3254245630 9647.0  
## + Personal 1 89859170 3284156909 9652.7  
## + Apps 1 76760285 3297255795 9655.2  
## + Accept 1 40515181 3333500898 9662.0  
## + Top10perc 1 19667478 3354348601 9665.9  
## + Top25perc 1 17264371 3356751709 9666.3  
## + Terminal 1 10921058 3363095022 9667.5  
## <none> 3374016079 9667.5  
## + Books 1 9703490 3364312590 9667.7  
## + PhD 1 7507701 3366508378 9668.1  
##   
## Step: AIC=9572.62  
## Outstate ~ Room.Board + perc.alumni + Expend + Private  
##   
## Df Sum of Sq RSS AIC  
## + Terminal 1 170639198 2717423838 9536.7  
## + PhD 1 160414885 2727648151 9539.0  
## + Grad.Rate 1 133557113 2754505923 9545.1  
## + Top25perc 1 71841644 2816221392 9558.9  
## + Top10perc 1 59011965 2829051071 9561.8  
## + Accept 1 37658886 2850404150 9566.4  
## + Personal 1 27767397 2860295639 9568.6  
## + S.F.Ratio 1 21368075 2866694961 9570.0  
## + Apps 1 11423432 2876639604 9572.1  
## <none> 2888063036 9572.6  
## + P.Undergrad 1 6015792 2882047244 9573.3  
## + Books 1 4756085 2883306951 9573.6  
## + Enroll 1 4166454 2883896582 9573.7  
## + F.Undergrad 1 280079 2887782957 9574.6  
##   
## Step: AIC=9536.68  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal  
##   
## Df Sum of Sq RSS AIC  
## + Grad.Rate 1 103124087 2614299751 9514.6  
## + S.F.Ratio 1 33278288 2684145550 9531.0  
## + Personal 1 32690397 2684733440 9531.1  
## + Top10perc 1 20164054 2697259783 9534.0  
## + Top25perc 1 19571296 2697852542 9534.2  
## + PhD 1 14375754 2703048084 9535.4  
## + P.Undergrad 1 14340337 2703083500 9535.4  
## + Accept 1 12472280 2704951558 9535.8  
## + Books 1 12182308 2705241530 9535.9  
## <none> 2717423838 9536.7  
## + F.Undergrad 1 6030101 2711393737 9537.3  
## + Apps 1 1704324 2715719514 9538.3  
## + Enroll 1 522301 2716901537 9538.6  
##   
## Step: AIC=9514.57  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate  
##   
## Df Sum of Sq RSS AIC  
## + S.F.Ratio 1 38414524 2575885227 9507.4  
## + Personal 1 21701648 2592598103 9511.4  
## + F.Undergrad 1 11317331 2602982420 9513.9  
## + Books 1 11007421 2603292330 9513.9  
## + PhD 1 8555785 2605743966 9514.5  
## <none> 2614299751 9514.6  
## + P.Undergrad 1 5003337 2609296414 9515.4  
## + Enroll 1 3942508 2610357243 9515.6  
## + Accept 1 3474593 2610825158 9515.7  
## + Top10perc 1 3264331 2611035420 9515.8  
## + Top25perc 1 2925135 2611374616 9515.9  
## + Apps 1 663933 2613635817 9516.4  
##   
## Step: AIC=9507.35  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio  
##   
## Df Sum of Sq RSS AIC  
## + Personal 1 25568232 2550316995 9503.1  
## + Books 1 10505967 2565379260 9506.8  
## + PhD 1 10079088 2565806139 9506.9  
## <none> 2575885227 9507.4  
## + F.Undergrad 1 7233313 2568651914 9507.6  
## + Accept 1 6191048 2569694179 9507.9  
## + P.Undergrad 1 4112236 2571772991 9508.4  
## + Top10perc 1 2367002 2573518225 9508.8  
## + Top25perc 1 2110307 2573774920 9508.8  
## + Enroll 1 1822652 2574062575 9508.9  
## + Apps 1 23217 2575862010 9509.3  
##   
## Step: AIC=9503.14  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal  
##   
## Df Sum of Sq RSS AIC  
## + PhD 1 10792340 2539524656 9502.5  
## + Accept 1 9470761 2540846234 9502.8  
## <none> 2550316995 9503.1  
## + Books 1 5267475 2545049521 9503.8  
## + Top10perc 1 3489116 2546827879 9504.3  
## + Top25perc 1 3098964 2547218031 9504.4  
## + F.Undergrad 1 2879037 2547437959 9504.4  
## + P.Undergrad 1 885263 2549431732 9504.9  
## + Enroll 1 233661 2550083334 9505.1  
## + Apps 1 217423 2550099573 9505.1  
##   
## Step: AIC=9502.49  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal + PhD  
##   
## Df Sum of Sq RSS AIC  
## + Accept 1 8422450 2531102206 9502.4  
## <none> 2539524656 9502.5  
## + Books 1 4465805 2535058851 9503.4  
## + F.Undergrad 1 3586261 2535938395 9503.6  
## + Top10perc 1 1396878 2538127777 9504.2  
## + Top25perc 1 1311941 2538212715 9504.2  
## + P.Undergrad 1 1077484 2538447171 9504.2  
## + Enroll 1 487904 2539036752 9504.4  
## + Apps 1 90166 2539434490 9504.5  
##   
## Step: AIC=9502.42  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal + PhD + Accept  
##   
## Df Sum of Sq RSS AIC  
## + F.Undergrad 1 52360778 2478741428 9491.4  
## + Enroll 1 40084255 2491017950 9494.5  
## + Apps 1 35390033 2495712172 9495.7  
## <none> 2531102206 9502.4  
## + Books 1 5476600 2525625606 9503.1  
## + P.Undergrad 1 3536188 2527566018 9503.6  
## + Top10perc 1 800496 2530301709 9504.2  
## + Top25perc 1 549028 2530553178 9504.3  
##   
## Step: AIC=9491.4  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal + PhD + Accept + F.Undergrad  
##   
## Df Sum of Sq RSS AIC  
## + Apps 1 37796231 2440945197 9483.8  
## <none> 2478741428 9491.4  
## + Books 1 4671620 2474069808 9492.2  
## + Top10perc 1 4559284 2474182144 9492.3  
## + Top25perc 1 3425837 2475315591 9492.5  
## + Enroll 1 753332 2477988096 9493.2  
## + P.Undergrad 1 66969 2478674459 9493.4  
##   
## Step: AIC=9483.83  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal + PhD + Accept + F.Undergrad +   
## Apps  
##   
## Df Sum of Sq RSS AIC  
## + Top10perc 1 21004761 2419940436 9480.4  
## + Top25perc 1 10331902 2430613295 9483.2  
## <none> 2440945197 9483.8  
## + Books 1 3930656 2437014541 9484.8  
## + Enroll 1 2184701 2438760496 9485.3  
## + P.Undergrad 1 53149 2440892048 9485.8  
##   
## Step: AIC=9480.44  
## Outstate ~ Room.Board + perc.alumni + Expend + Private + Terminal +   
## Grad.Rate + S.F.Ratio + Personal + PhD + Accept + F.Undergrad +   
## Apps + Top10perc  
##   
## Df Sum of Sq RSS AIC  
## <none> 2419940436 9480.4  
## + Books 1 5857894 2414082542 9480.9  
## + Enroll 1 4248300 2415692136 9481.3  
## + P.Undergrad 1 1053513 2418886923 9482.2  
## + Top25perc 1 433009 2419507426 9482.3

summary(stepwise\_model)

##   
## Call:  
## lm(formula = Outstate ~ Room.Board + perc.alumni + Expend + Private +   
## Terminal + Grad.Rate + S.F.Ratio + Personal + PhD + Accept +   
## F.Undergrad + Apps + Top10perc, data = train\_set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6432 -1304 -8 1270 10081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.597e+03 8.313e+02 -1.921 0.05516 .   
## Room.Board 9.796e-01 9.693e-02 10.106 < 2e-16 \*\*\*  
## perc.alumni 4.120e+01 8.531e+00 4.830 1.73e-06 \*\*\*  
## Expend 1.762e-01 2.440e-02 7.219 1.57e-12 \*\*\*  
## PrivateYes 2.094e+03 2.786e+02 7.514 2.05e-13 \*\*\*  
## Terminal 2.815e+01 1.133e+01 2.484 0.01327 \*   
## Grad.Rate 2.394e+01 6.044e+00 3.961 8.33e-05 \*\*\*  
## S.F.Ratio -8.282e+01 2.930e+01 -2.827 0.00486 \*\*   
## Personal -2.396e-01 1.325e-01 -1.809 0.07101 .   
## PhD 1.153e+01 1.046e+01 1.102 0.27087   
## Accept 6.631e-01 1.270e-01 5.220 2.46e-07 \*\*\*  
## F.Undergrad -1.590e-01 3.900e-02 -4.078 5.14e-05 \*\*\*  
## Apps -2.631e-01 7.121e-02 -3.695 0.00024 \*\*\*  
## Top10perc 1.726e+01 7.507e+00 2.299 0.02183 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1993 on 609 degrees of freedom  
## Multiple R-squared: 0.7622, Adjusted R-squared: 0.7571   
## F-statistic: 150.1 on 13 and 609 DF, p-value: < 2.2e-16

library(mgcv)

## Loading required package: nlme

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.

gam\_model <- gam(Outstate ~ s(Terminal) + s(Accept) + s(Room.Board) + s(S.F.Ratio) + s(Apps) + s(Expend) + s(Personal) + s(Top10perc) + Private + s(PhD), data = train\_set)  
  
test\_predictions <- predict(gam\_model, newdata = test\_set)  
  
  
r\_squared <- 1 - sum((test\_set$Outstate - test\_predictions)^2) / sum((test\_set$Outstate - mean(test\_set$Outstate))^2)  
  
r\_squared

## [1] 0.7734406

summary(gam\_model)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## Outstate ~ s(Terminal) + s(Accept) + s(Room.Board) + s(S.F.Ratio) +   
## s(Apps) + s(Expend) + s(Personal) + s(Top10perc) + Private +   
## s(PhD)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8253.5 208.2 39.63 <2e-16 \*\*\*  
## PrivateYes 3000.0 266.7 11.25 <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(Terminal) 1.000 1.000 1.764 0.18464   
## s(Accept) 6.234 7.145 2.241 0.02760 \*   
## s(Room.Board) 1.000 1.000 82.325 < 2e-16 \*\*\*  
## s(S.F.Ratio) 6.590 7.717 3.099 0.00206 \*\*   
## s(Apps) 6.085 7.039 1.953 0.05829 .   
## s(Expend) 4.771 5.896 19.874 < 2e-16 \*\*\*  
## s(Personal) 2.882 3.642 5.039 0.00109 \*\*   
## s(Top10perc) 1.000 1.000 3.153 0.07629 .   
## s(PhD) 6.318 7.409 2.089 0.04459 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.787 Deviance explained = 80%  
## GCV = 3.7014e+06 Scale est. = 3.4763e+06 n = 623

The model has a high R\_sqared which is good.

ALl varibles seem to be non linnear except Terminal, Room.Board, and Top10percent.