Assignment #5

Question 1

Use College data set.

```
library(ISLR)
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

a) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

Split the data into a training set and test set

```
# Split data with the ratio 50:50
set.seed(1)
split<- sample(c(rep(0, 0.5 * nrow(College)), rep(1, 0.5 * nrow(College))))
training <- College[split == 0, ]
test <- College[split == 1, ]
```

Perform forward stepwise selection on the training set

```
library(leaps)
# Perform forward stepwise selection to choose the best model
forward <- regsubsets(Outstate ~., data = training, nvmax = 17, method = "for
ward")
summary(forward)
```

```
## Subset selection object
## Call: regsubsets.formula(Outstate ~ ., data = training, nvmax = 17, 
##     method = "forward")
## 17 Variables  (and intercept)
##             Forced in Forced out
## PrivateYes      FALSE      FALSE
## Apps            FALSE      FALSE
## Accept          FALSE      FALSE
## Enroll          FALSE      FALSE
## Top10perc       FALSE      FALSE
## Top25perc       FALSE      FALSE
## F.Undergrad     FALSE      FALSE
## P.Undergrad     FALSE      FALSE
## Room.Board      FALSE      FALSE
## Books           FALSE      FALSE
## Personal        FALSE      FALSE
## PhD             FALSE      FALSE
## Terminal        FALSE      FALSE
## S.F.Ratio       FALSE      FALSE
## perc.alumni     FALSE      FALSE
## Expend          FALSE      FALSE
## Grad.Rate       FALSE      FALSE
## 1 subsets of each size up to 17
## Selection Algorithm: forward
##           PrivateYes Apps Accept Enroll Top10perc Top25perc F.Undergrad
## 1  ( 1 )  " "        " "  " "    " "    " "       " "       " "        
## 2  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 3  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 4  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 5  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 6  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 7  ( 1 )  "*"        " "  " "    " "    " "       " "       " "        
## 8  ( 1 )  "*"        " "  " "    " "    " "       "*"       " "        
## 9  ( 1 )  "*"        " "  " "    " "    " "       "*"       " "        
## 10  ( 1 ) "*"        " "  "*"    " "    " "       "*"       " "        
## 11  ( 1 ) "*"        " "  "*"    " "    " "       "*"       "*"        
## 12  ( 1 ) "*"        "*"  "*"    " "    " "       "*"       "*"        
## 13  ( 1 ) "*"        "*"  "*"    " "    "*"       "*"       "*"        
## 14  ( 1 ) "*"        "*"  "*"    " "    "*"       "*"       "*"        
## 15  ( 1 ) "*"        "*"  "*"    " "    "*"       "*"       "*"        
## 16  ( 1 ) "*"        "*"  "*"    "*"    "*"       "*"       "*"        
## 17  ( 1 ) "*"        "*"  "*"    "*"    "*"       "*"       "*"        
##           P.Undergrad Room.Board Books Personal PhD Terminal S.F.Ratio
## 1  ( 1 )  " "         " "        " "   " "      " " " "      " "      
## 2  ( 1 )  " "         " "        " "   " "      " " " "      " "      
## 3  ( 1 )  " "         "*"        " "   " "      " " " "      " "      
## 4  ( 1 )  " "         "*"        " "   " "      " " " "      " "      
## 5  ( 1 )  " "         "*"        " "   " "      " " " "      " "      
## 6  ( 1 )  " "         "*"        " "   " "      "*" " "      " "      
## 7  ( 1 )  " "         "*"        " "   "*"      "*" " "      " "      
```

```
## 8  ( 1 )  " "         "*"        " "   "*"      "*" " "       " "
## 9  ( 1 )  " "         "*"        " "   "*"      "*" " "       "*"
## 10  ( 1 ) " "         "*"        " "   "*"      "*" " "       "*"
## 11  ( 1 ) " "         "*"        " "   "*"      "*" " "       "*"
## 12  ( 1 ) " "         "*"        " "   "*"      "*" " "       "*"
## 13  ( 1 ) " "         "*"        " "   "*"      "*" " "       "*"
## 14  ( 1 ) "*"         "*"        " "   "*"      "*" " "       "*"
## 15  ( 1 ) "*"         "*"        " "   "*"      "*" "*"       "*"
## 16  ( 1 ) "*"         "*"        " "   "*"      "*" "*"       "*"
## 17  ( 1 ) "*"         "*"        "*"   "*"      "*" "*"       "*"
##           perc.alumni Expend Grad.Rate
## 1  ( 1 )  " "         "*"    " "
## 2  ( 1 )  " "         "*"    " "
## 3  ( 1 )  " "         "*"    " "
## 4  ( 1 )  "*"         "*"    " "
## 5  ( 1 )  "*"         "*"    "*"
## 6  ( 1 )  "*"         "*"    "*"
## 7  ( 1 )  "*"         "*"    "*"
## 8  ( 1 )  "*"         "*"    "*"
## 9  ( 1 )  "*"         "*"    "*"
## 10  ( 1 ) "*"         "*"    "*"
## 11  ( 1 ) "*"         "*"    "*"
## 12  ( 1 ) "*"         "*"    "*"
## 13  ( 1 ) "*"         "*"    "*"
## 14  ( 1 ) "*"         "*"    "*"
## 15  ( 1 ) "*"         "*"    "*"
## 16  ( 1 ) "*"         "*"    "*"
## 17  ( 1 ) "*"         "*"    "*"
```

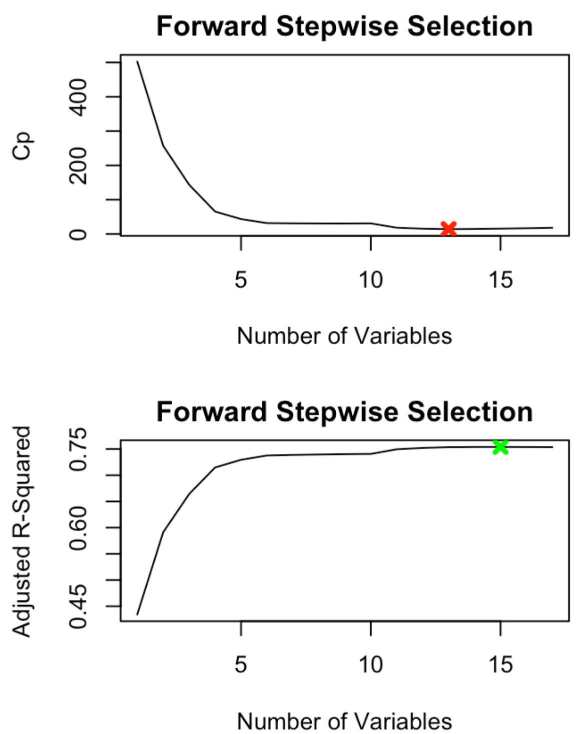
Select the predictors using C_p, BIC, and Adjusted R^2 to get a satisfactory model

```
# Obtain Cp, BIC, and adjusted R2
sum <- summary(forward)
test.error <- data.frame(
  Cp = which.min(sum$cp),
  BIC = which.min(sum$bic),
  Adj.R2 = which.max(sum$adjr2)
)
print(test.error)

##   Cp BIC Adj.R2
## 1 13   6     15
```

We found that:

- The best model using the lowest C_p is the model includes 13 predictors.
- The best model using the lowest BIC is the model includes 6 predictors.
- The best model using the highest Adjusted R^2 is the model includes 15 predictors.

```
par(mfrow=c(2,2))
plot(sum$cp, xlab="Number of Variables", ylab="Cp", pch=20, type="l", main =
"Forward Stepwise Selection")
points(13, sum$cp[13], pch=4, col="red", lwd=3)
plot(sum$bic, xlab="Number of Variables", ylab="BIC", pch=20, type="l", main
= "Forward Stepwise Selection")
points(6, sum$bic[6], pch=4, col="blue", lwd=3)
plot(sum$adjr2, xlab="Number of Variables", ylab="Adjusted R-Squared", pch=20
, type="l", main  = "Forward Stepwise Selection")
points(15, sum$adjr2[15], pch=4, col="green", lwd=3)
```



We found that with C_p, BIC, and adjusted R^2, the best model contains set of 13, 6, and 15, respectively. Therefore, we will choose the model using the lowest BIC to obtain the coefficients of the best model.

A satisfactory model that uses just a subset of the predictors

```
# The coefficients of the best model
best.mod <- coef(forward,6)
print(best.mod)

##   (Intercept)    PrivateYes    Room.Board           PhD   perc.alumni
## -3456.7207973  2590.2527838     0.9790685    28.2946510    62.6467562
##        Expend     Grad.Rate
##     0.2288976    31.7893157
```

The best subset contains set of 6 predictors as follows:

1) Private
2) Room.Board
3) PhD
4) perc.alumni
5) Expend
6) Grad.Rate

b. Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results and explain your findings.

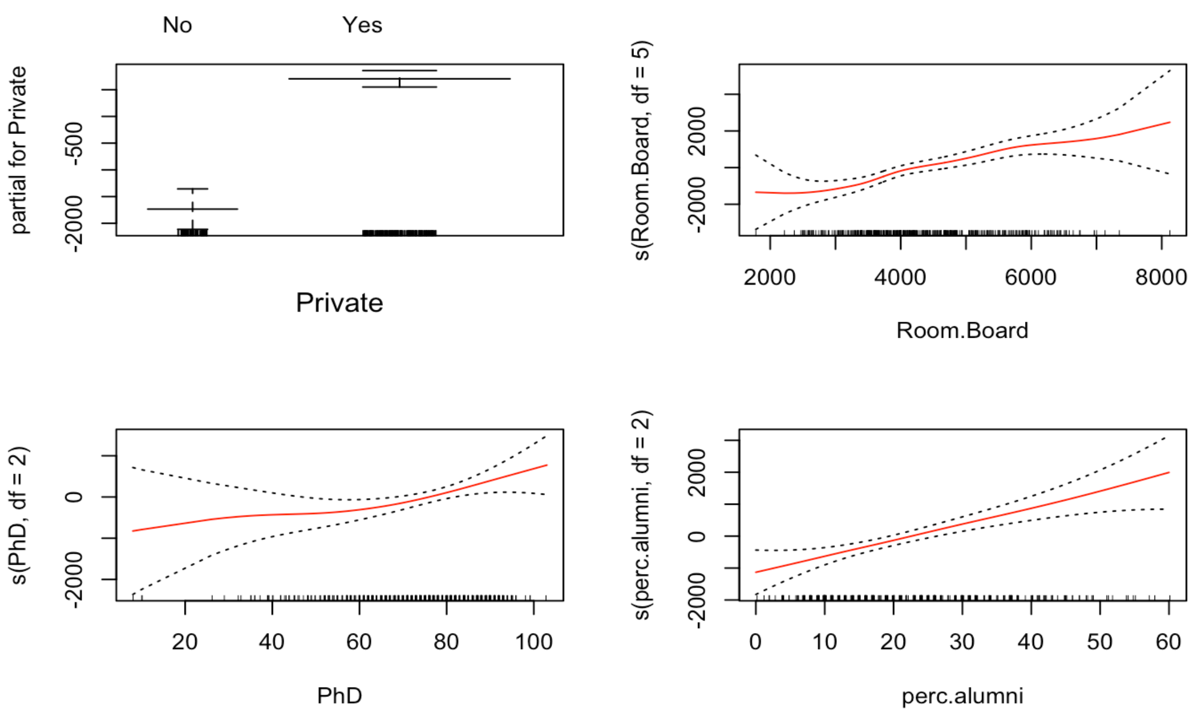
Fit a GAM on the training data

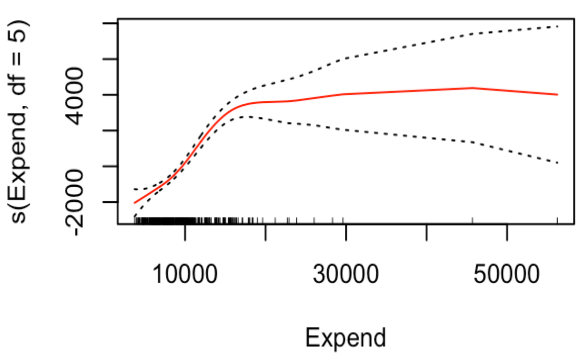
We specify that the functions of the predictors PhD, perc.alumni, and Grad.Rate should have 2 degrees of freedom and the functions of Expend and Room.Board should have 5 degrees of freedom. We cannot fit a smoothing spline to the Private variable since it is a factor.

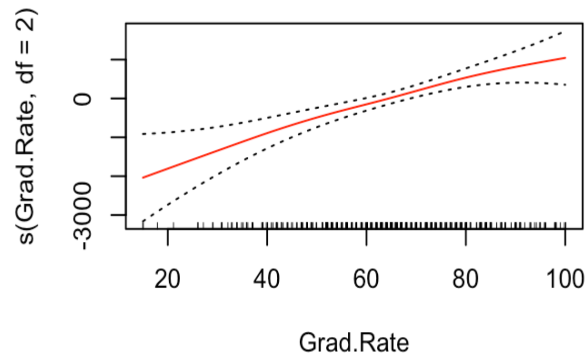
```
library(gam)
gam.fit <- gam(Outstate ~ Private + s(Room.Board, df = 5) + s(PhD, df = 2) +
s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate, df = 2), data = tra
ining)
```

Plot the results

```
par(mfrow = c(2,2))
plot(gam.fit, se = T, col = 'red
```







We have used the GAM function to fit smoothing spline with 2 degrees of freedom to some predictors and 5 degrees of freedom to Expend and Room.Board. Private, one of the predictors selected by the forward stepwise selection, is a dummy variable, so it is not fit to a smoothing spline.

Among these plots, the Expend function plot does not look linear; therefore, this may show some evidence of nonlinear relationships in the data.

(c) Evaluate the model obtained on the test set, and explain the results obtained.

```
# Evaluate the model using MSE
y.pred <- predict(gam.fit, newdata = test)
y.test <- test$Outstate
test.err <- mean((y.test - y.pred)^2)
sprintf('%s is %f', 'The test error for the GAM model', test.err)

## [1] "The test error for the GAM model is 3386481.289178"



# Evaluate the model using R-Squared
TSS <- sum((test$Outstate - mean(test$Outstate))^2)
RSS <- sum((y.pred - test$Outstate)^2)
R2 <- 1 - RSS/TSS
sprintf('%s is %f', 'The R-Squared for the GAM model', R2)

## [1] "The R-Squared for the GAM model is 0.790220"
```

From the results, we can obtain that the test error of this model is **3386481.289178**, and the R-squared is **0.79** meaning that 79% of the variation in the out-of-state tuition variable is explained by these 6 predictors.

d) For which variables, if any, is there evidence of a non-linear relationship with the response?

```
summary(gam.fit)

##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 5) + s(PhD,
##     df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate,
##     df = 2), data = training)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -7180.64 -1119.35    31.84  1224.01  7657.37
##
## (Dispersion Parameter for gaussian family taken to be 3612061)
##
##     Null Deviance: 6278511356 on 387 degrees of freedom
## Residual Deviance: 1336461815 on 369.9998 degrees of freedom
## AIC: 6979.384
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                         Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private                  1 1774368633 1774368633 491.234 < 2.2e-16 ***
## s(Room.Board, df = 5)    1 1223128276 1223128276 338.623 < 2.2e-16 ***
## s(PhD, df = 2)           1  317151454  317151454  87.803 < 2.2e-16 ***
## s(perc.alumni, df = 2)   1  340707261  340707261  94.325 < 2.2e-16 ***
## s(Expend, df = 5)        1  453399132  453399132 125.524 < 2.2e-16 ***
## s(Grad.Rate, df = 2)     1   87073215   87073215  24.106  1.37e-06 ***
## Residuals              370 1336461815    3612061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                        Npar Df  Npar F     Pr(F)
## (Intercept)
## Private
## s(Room.Board, df = 5)        4  0.9548    0.4323
## s(PhD, df = 2)               1  2.8207    0.0939 .
## s(perc.alumni, df = 2)       1  0.2242    0.6361
## s(Expend, df = 5)            4 16.4362 2.092e-12 ***
## s(Grad.Rate, df = 2)         1  1.6450    0.2004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA for Nonparametric effect test, there is a strong evidence that the variable **Expend** indicates non-linear relationship with the response, out-of-state tuition, with a significance level of 0.001. In fact, there is some evidence of a nonlinear effect of **PhD,** with a significance level of 0.1.

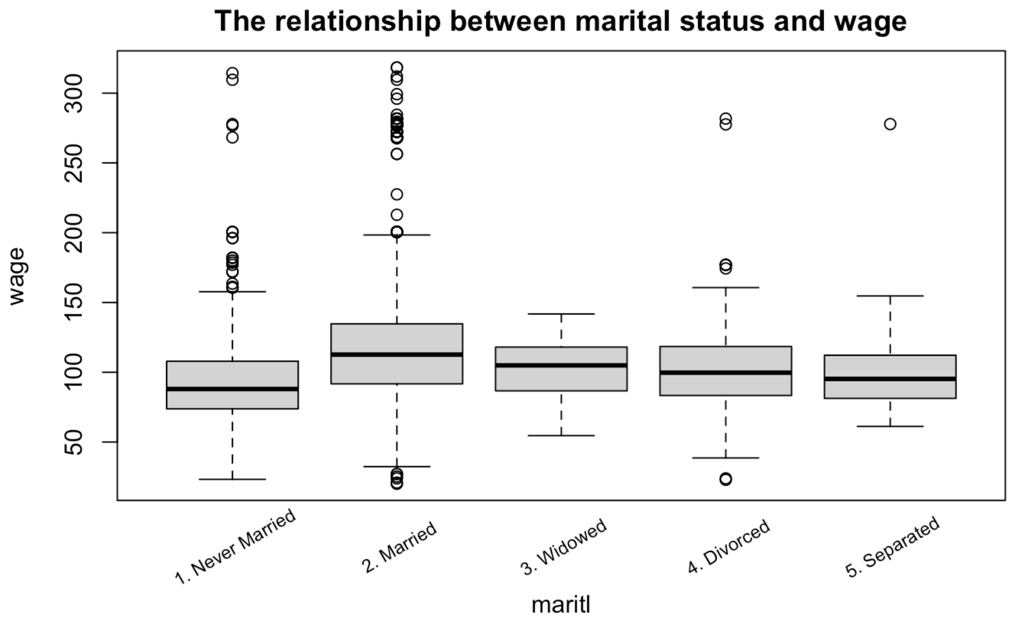
Question 2

The Wage data set contains a number of other features not explored in this chapter, such as marital status (maritl), job class (jobclass), and others. Explore the relationships between some of these other predictors and wage and use non-linear fitting techniques in order to fit flexible models to the data. Create plots of the results obtained and write a summary of your findings.

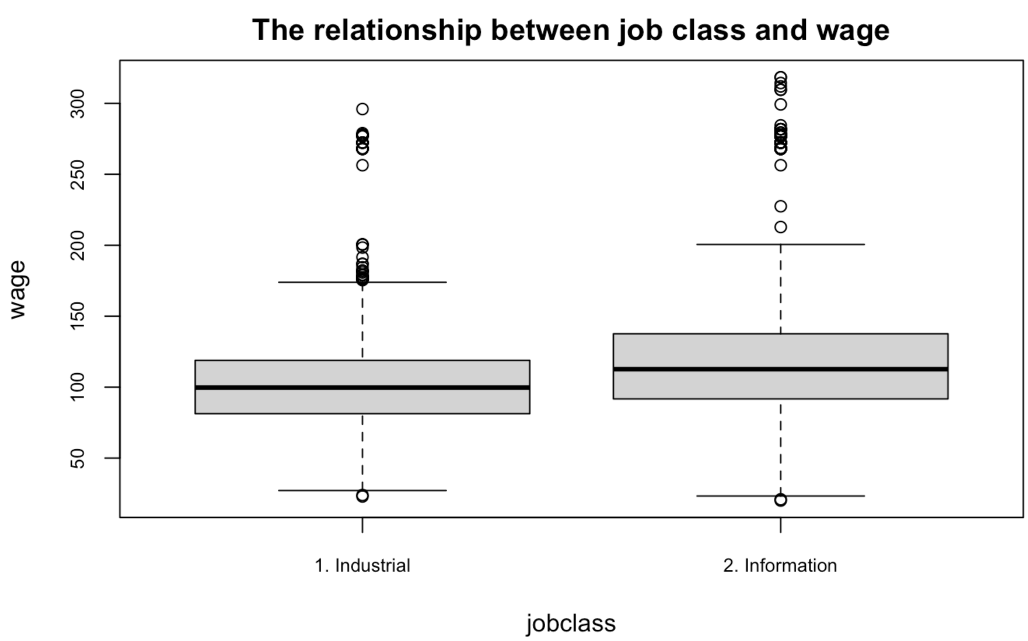
```
# Load Wage data set
library(ISLR)
str(Wage)

## 'data.frame':    3000 obs. of  11 variables:
##  $ year      : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
##  $ age       : int  18 24 45 43 50 54 44 30 41 52 ...
##  $ maritl    : Factor w/ 5 levels "1. Never Married",..: 1 1 2 2 4 2 2 1 1
2 ...
##  $ race      : Factor w/ 4 levels "1. White","2. Black",..: 1 1 1 3 1 1 4
3 2 1 ...
##  $ education : Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 .
..
##  $ region    : Factor w/ 9 levels "1. New England",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ jobclass  : Factor w/ 2 levels "1. Industrial",..: 1 2 1 2 2 2 1 2 2 2
...
##  $ health    : Factor w/ 2 levels "1. <=Good","2. >=Very Good": 1 2 1 2 1
2 2 1 2 2 ...
##  $ health_ins: Factor w/ 2 levels "1. Yes","2. No": 2 2 1 1 1 1 1 1 1 1 ..
.
##  $ logwage   : num  4.32 4.26 4.88 5.04 4.32 ...
##  $ wage      : num  75 70.5 131 154.7 75 ...
```

In the class, we explored the relationships between year, age, and education and wage. We, therefore, are going to explore the relationships between some of following predictors and wage: *marital status and job class.*

```
# Explore the relationships between some of these other predictors and wage
# Marital status and wage
boxplot(wage~maritl, data = Wage, pars  =  list(xaxt = "n"))
text(1:5, par("usr")[3] - 20, labels = levels(Wage$maritl), cex = 0.75, srt =
30, pos = 1, xpd = TRUE)
title('The relationship between marital status and wage')
```




```
# Job class and wage
boxplot(wage~jobclass, data = Wage, cex.axis=0.75)
title('The relationship between job class and wage')
```



From the plots, we can observe that:

- Married people seem to have higher wages.
- People doing informational jobs seem to have higher wages.

Now, we will fit GAM models, non-linear fitting techniques, to fit flexible models to the data.

As it was mentioned in class that GAMs allow us to fit non-linear functions to each variable; therefore, we will use this non-linear fitting technique to fit models to the data. Therefore, we use the linear regression of year term. We specify that the function of age term should have 5 degrees of freedom as well as adding some qualitative features to models such as education, marital status, and job class.

```
library(gam)
# Fit flexible GAM models
gam.fit1 <- gam(wage ~ year + s(age, df = 5), data = Wage)
gam.fit2 <- gam(wage ~ year + s(age, df = 5) + education, data = Wage)
gam.fit3 <- gam(wage ~ year + s(age, df = 5) + education + maritl, data = Wag
e)
gam.fit4 <- gam(wage ~ year + s(age, df = 5) + education + jobclass, data = W
age)
gam.fit5 <- gam(wage ~ year + s(age, df = 5) + education + maritl + jobclass,
data = Wage)

# ANOVA test for the best model
anova(gam.fit1, gam.fit2, gam.fit3, gam.fit4,gam.fit5)

## Analysis of Deviance Table
##
## Model 1: wage ~ year + s(age, df = 5)
## Model 2: wage ~ year + s(age, df = 5) + education
## Model 3: wage ~ year + s(age, df = 5) + education + maritl
## Model 4: wage ~ year + s(age, df = 5) + education + jobclass
## Model 5: wage ~ year + s(age, df = 5) + education + maritl + jobclass
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2993    4741791
## 2      2989    3693842  4  1047950 < 2.2e-16 ***
## 3      2985    3599643  4    94198 3.804e-16 ***
## 4      2988    3681289 -3   -81646 1.172e-14 ***
## 5      2984    3585383  4    95906 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

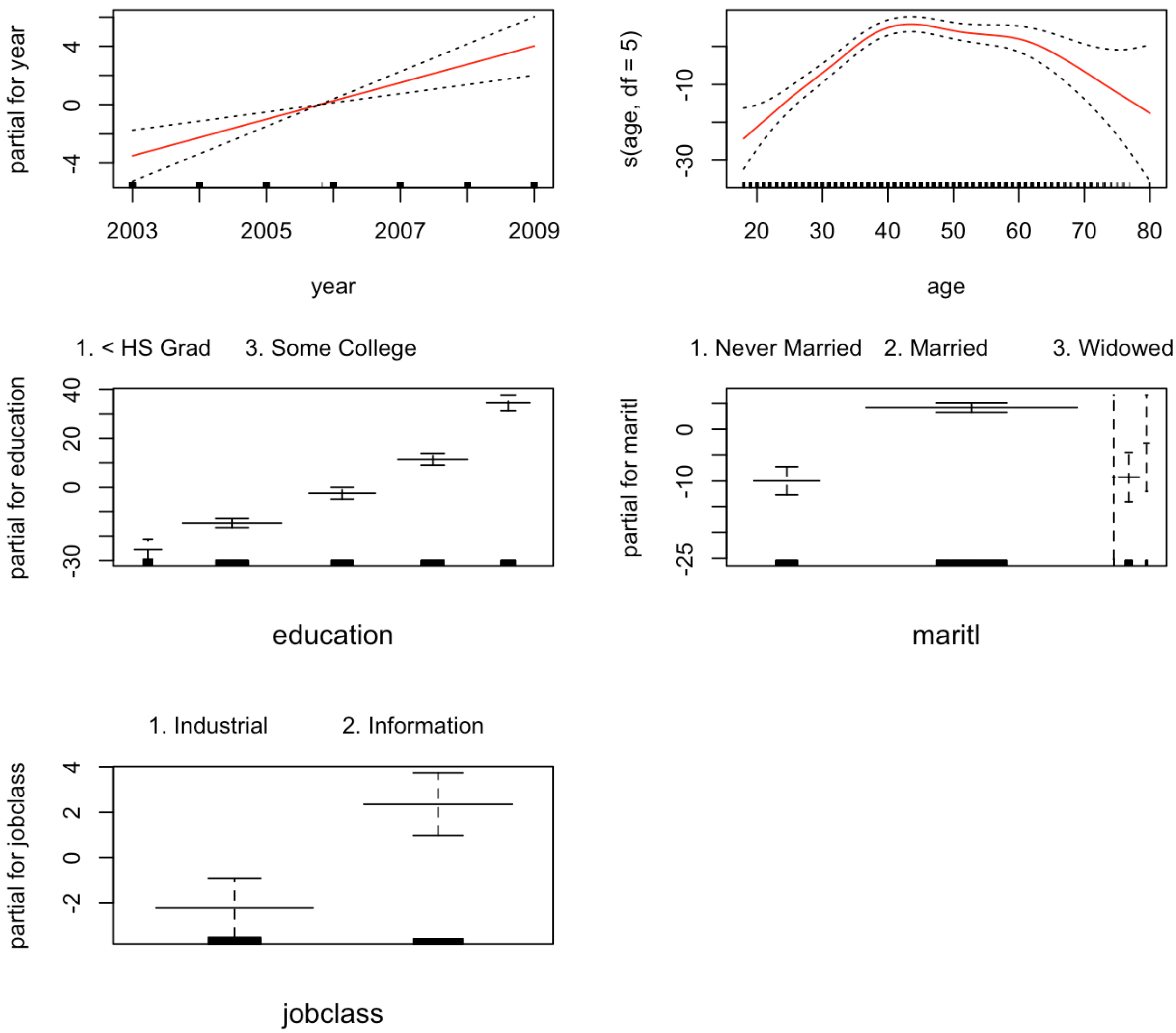
From the results, we find that there is compelling evidence that the GAM model with follow ing predictors fits the best.

- year
- age
- education
- maritl: Marital status
- jobclass: job class

Therefore, the model 5 will be the best model to the Wage data.

Plot the results

```
# plot the results of the selected model:
par(mfrow = c(2, 2))
plot(gam.fit5, se = T, col = "red")
```



For the Wage data, plots of the relationship between each feature and the response, wage, in the fitted model are shown above. Each plot displays the fitted function and pointwise standard errors. The first function is linear regression in year, the second function is a smoothing spline with five degrees of freedom in age, and the last three functions are the qualitative variables education, marital status, and job class.

Conclusion:

Again, the plots show that people doing informational jobs and married people have higher wages. Apart from marital status and job class, the plots show positive relationships between year and wage as well as education and wage meaning that the wage increases when education and year increase. In addition, middle age people seem to have higher wages.