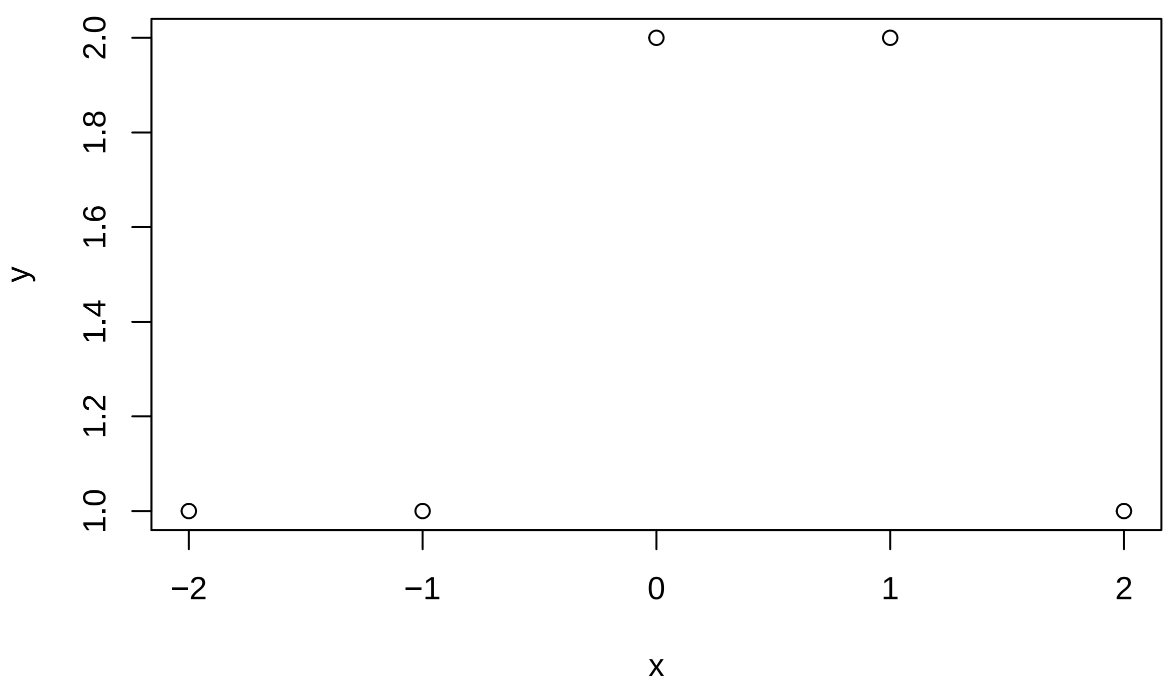
Stat 101C HW5

Junhyuk Jang

May 23, 2017

SID: 004 728 134 DIS: 2A

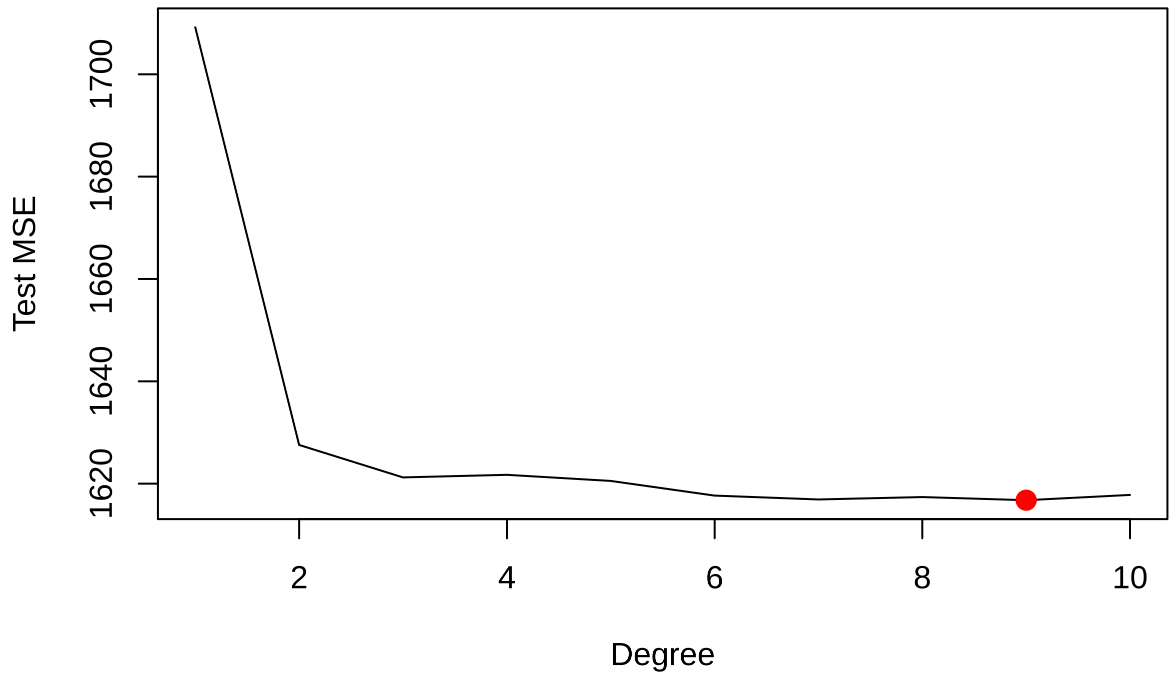
```
# 4
x = -2:2
y = c(1 + 0 + 0, # x = -2
      1 + 0 + 0, # x = -1
      1 + 1 + 0, # x = 0
      1 + (1-0) + 0, # x = 1
      1 + (1-1) + 0 # x =2
      )
plot(x,y)
```



```
# 1. y = 3-x between 1 and 2
# 2. y = 2 between 0 and 1
# 3. y = 1 between -2 and 0

#6.
#a)
set.seed(1)
Wage <- read.csv("~/Desktop/WageLec2.csv")
attach(Wage)
library(boot)
```

```
## Warning: package 'boot' was built under R version 3.2.5
```

```r
del <- rep(NA, 10)
for (i in 1:10) {
    fit <- glm(wage ~ poly(age, i), data = Wage)
    del[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(1:10, del, xlab = "Degree", ylab = "Test MSE", type = "l")
del_min <- which.min(del)
points(which.min(del), del[which.min(del)], col = "red", cex = 2, pch = 20)
```

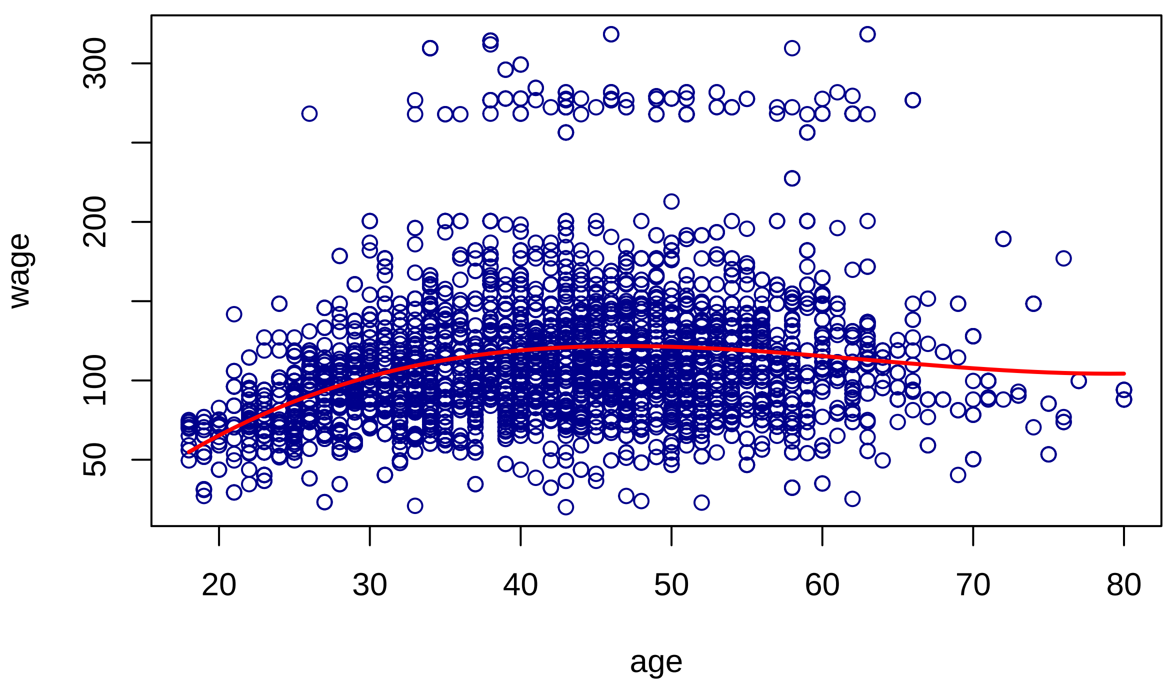



```r
# The polynomial degree 9 minimized the test MSE.
# It is require to test using ANOVA that whether M1 is sufficiently explain the data or
# we need more complex model to explain the data.

fit.1 = lm(wage~poly(age, 1), data=Wage)
fit.2 = lm(wage~poly(age, 2), data=Wage)
fit.3 = lm(wage~poly(age, 3), data=Wage)
fit.4 = lm(wage~poly(age, 4), data=Wage)
fit.5 = lm(wage~poly(age, 5), data=Wage)
fit.6 = lm(wage~poly(age, 6), data=Wage)
fit.7 = lm(wage~poly(age, 7), data=Wage)
fit.8 = lm(wage~poly(age, 8), data=Wage)
fit.9 = lm(wage~poly(age, 9), data=Wage)
fit.10 = lm(wage~poly(age, 10), data=Wage)
anova(fit.1, fit.2, fit.3, fit.4, fit.5, fit.6, fit.7, fit.8, fit.9, fit.10)
```

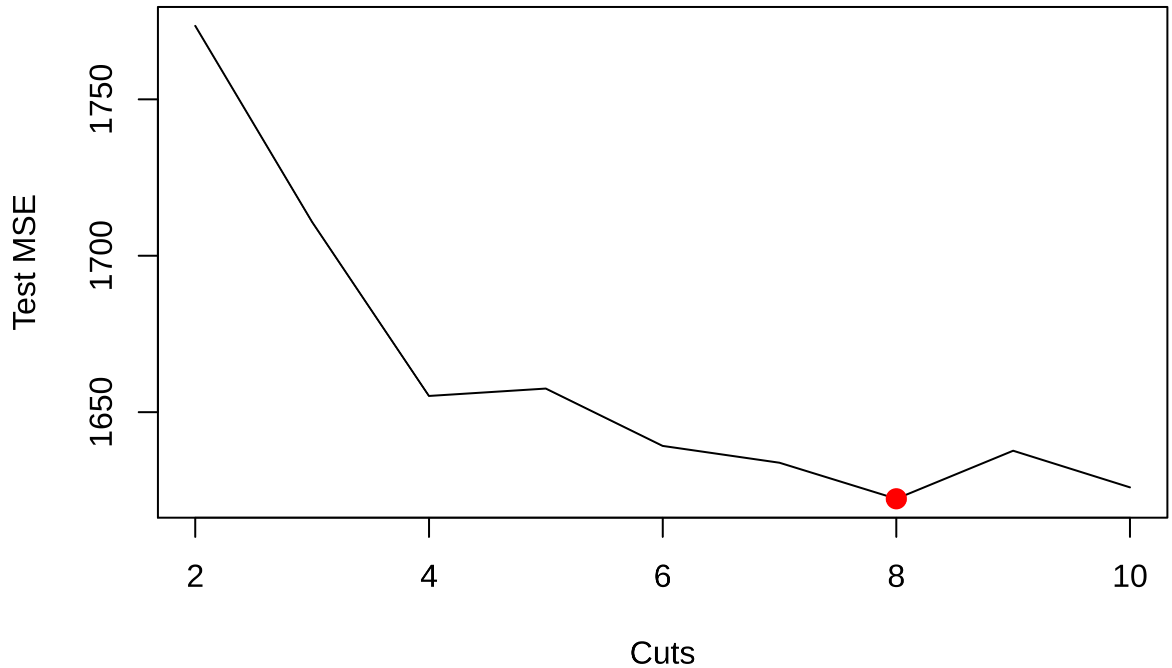
```
## Analysis of Variance Table
##
## Model  1: wage ~ poly(age, 1)
## Model  2: wage ~ poly(age, 2)
## Model  3: wage ~ poly(age, 3)
## Model  4: wage ~ poly(age, 4)
```

```
## Model  5: wage ~ poly(age, 5)
## Model  6: wage ~ poly(age, 6)
## Model  7: wage ~ poly(age, 7)
## Model  8: wage ~ poly(age, 8)
## Model  9: wage ~ poly(age, 9)
## Model 10: wage ~ poly(age, 10)
##    Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1    3998 6827523
## 2    3997 6503669  1    323854 200.6080 < 2.2e-16 ***
## 3    3996 6471970  1     31699  19.6353 9.624e-06 ***
## 4    3995 6469894  1      2076   1.2859  0.256881
## 5    3994 6457099  1     12795   7.9256  0.004898 **
## 6    3993 6452761  1      4339   2.6875  0.101220
## 7    3992 6446093  1      6668   4.1304  0.042186 *
## 8    3991 6446068  1        24   0.0151  0.902150
## 9    3990 6441046  1      5022   3.1109  0.077845 .
## 10   3989 6439693  1      1353   0.8381  0.360008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

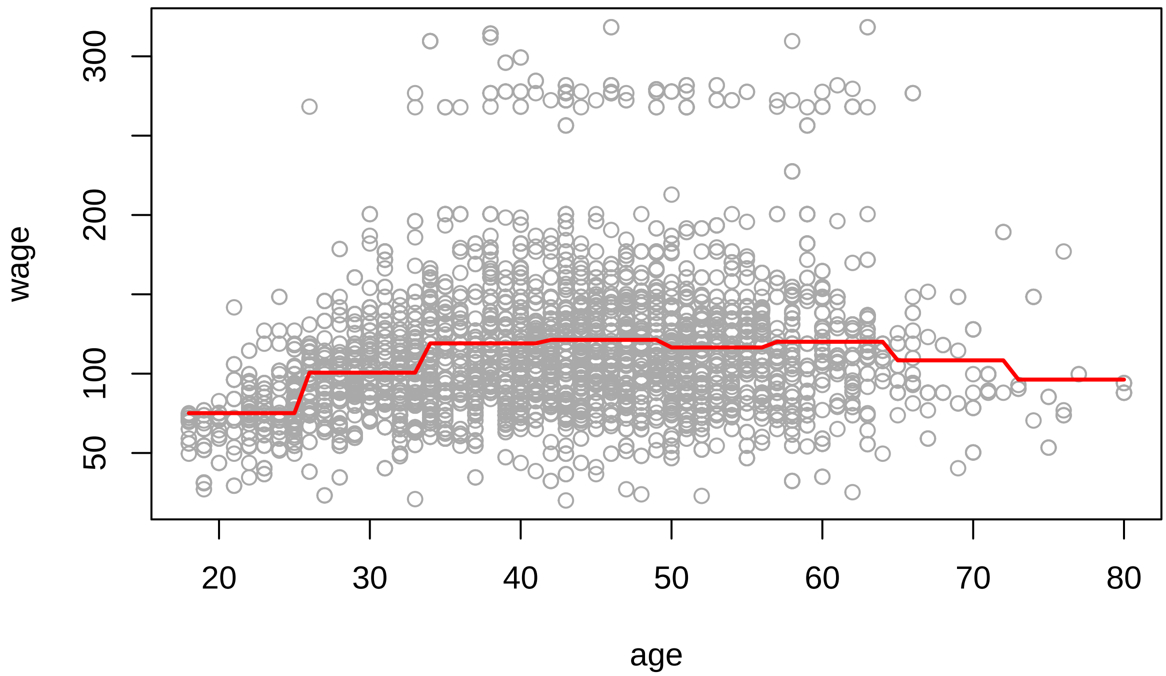
```r
# Anova comparison shows that more than 3 polynomial degree models are
# statistically insignificant with 0.001 significance level.

plot(wage ~ age, data = Wage, col = "darkblue")
agelims <- range(Wage$age)
age.grid <- seq(from = agelims[1], to = agelims[2])
fit <- lm(wage ~ poly(age, 3), data = Wage)
preds <- predict(fit, newdata = list(age = age.grid))
lines(age.grid, preds, col = "red", lwd = 2)
```



```r
# b)
cv <- rep(NA, 10)
for (i in 2:10) {
    Wage$age.cut <- cut(Wage$age, i)
    fit <- glm(wage ~ age.cut, data = Wage)
    cv[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(2:10, cv[-1], xlab = "Cuts", ylab = "Test MSE", type = "l")
min <- which.min(cv)
points(which.min(cv), cv[which.min(cv)], col = "red", cex = 2, pch = 20)
```




```r
# The plot shows that with 8 cuts we can mimize the Test MSE

plot(wage ~ age, data = Wage, col = "darkgrey")
age <- range(Wage$age)
grid <- seq(from = age[1], to = age[2])
fit <- glm(wage ~ cut(age, 8), data = Wage)
preds <- predict(fit, data.frame(age = grid))
lines(grid, preds, col = "red", lwd = 2)
```



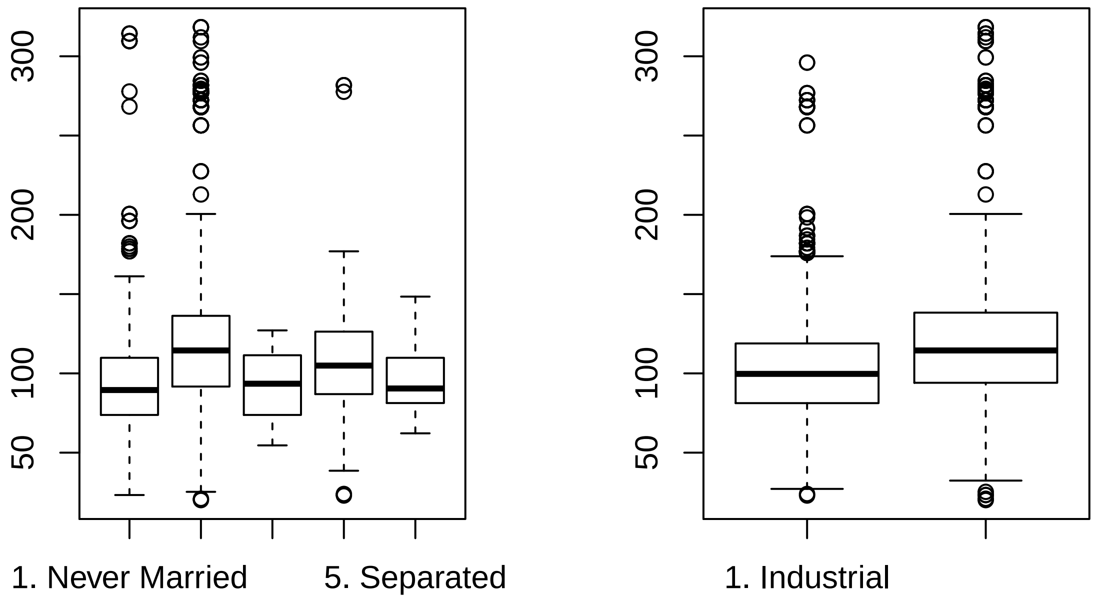

```r
# 7.
set.seed(1)
summary(Wage$maritl)
```

```
## 1. Never Married       2. Married       3. Widowed      4. Divorced
##               865             2762               18              294
##      5. Separated
##                61
```

```r
summary(Wage$jobclass)
```

```
##  1. Industrial 2. Information
##           2006           1994
```

```r
par(mfrow = c(1, 2))
plot(Wage$maritl, Wage$wage)
plot(Wage$jobclass, Wage$wage)
```




```r
# We can say that informational jobs has higher wage than industiral job on average.
# The plot shows that the married person make more money on average compare to the other
# groups

# install.packages("gam")
library(gam)
```

```
## Warning: package 'gam' was built under R version 3.2.5

## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.14-4
```

```r
fit0 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education, data = Wage)
fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, data = Wage)
fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl, data = Wage)
fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass + maritl, data = Wage)
anova(fit0, fit1, fit2, fit3)
```

```
## Analysis of Deviance Table
##
## Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
## Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
## Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
## Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
##     maritl
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    3987.1    4942702
## 2    3986.1    4909333  1    33368 1.461e-07 ***
## 3    3983.1    4845935  3    63399 2.325e-11 ***
```

```
## 4    3982.1    4807302  1    38632 1.541e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Based on the p-value model 3 is prefered than the others (Lowest p-value).

par(mfrow = c(3, 2))
plot(fit3, se = T, col = "blue")

# 10
# a)
library(leaps)
college <- read.csv("~/Desktop/CollegeLec2.csv")
attach(college)
```

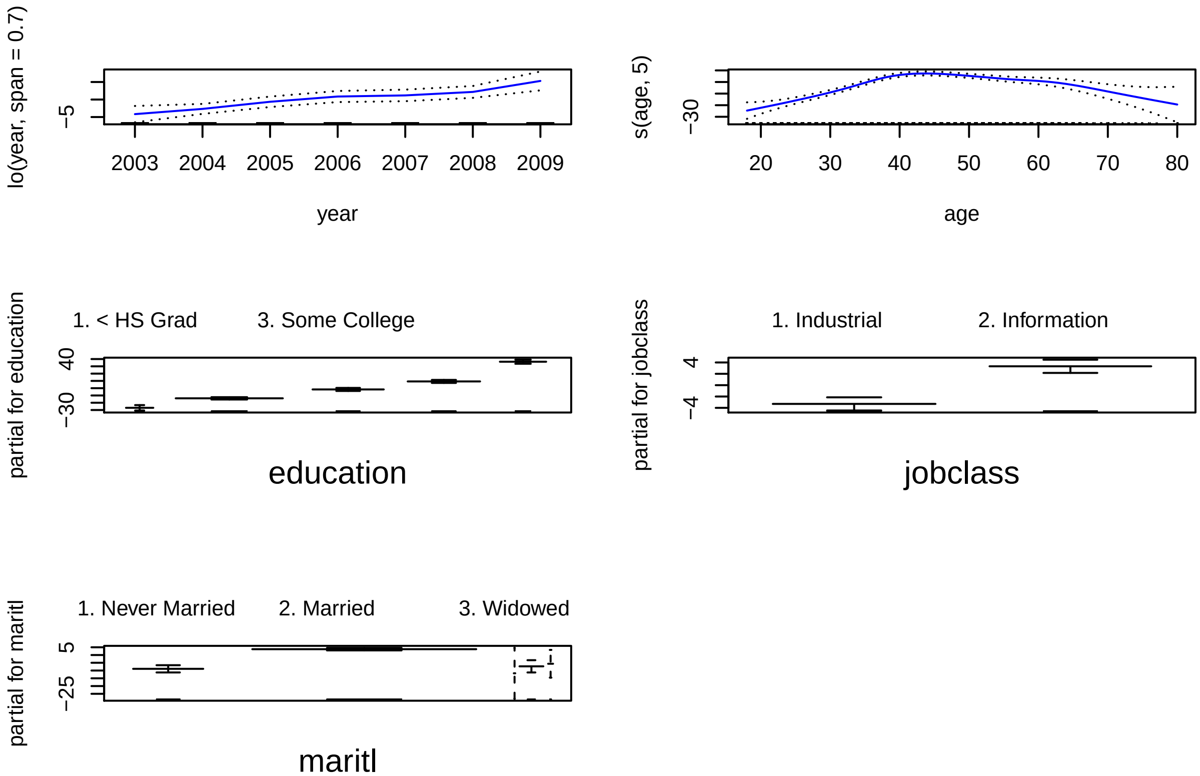
```
## The following object is masked from Wage:
##
##     X
```

```r
college <- college[,-1]

train <- sample(length(Outstate), length(Outstate) / 2)
test <- -train
college.train <- college[train, ]
college.test <- college[test, ]
fit <- regsubsets(Outstate ~ ., data = college.train, nvmax = 19, method = "forward")
```

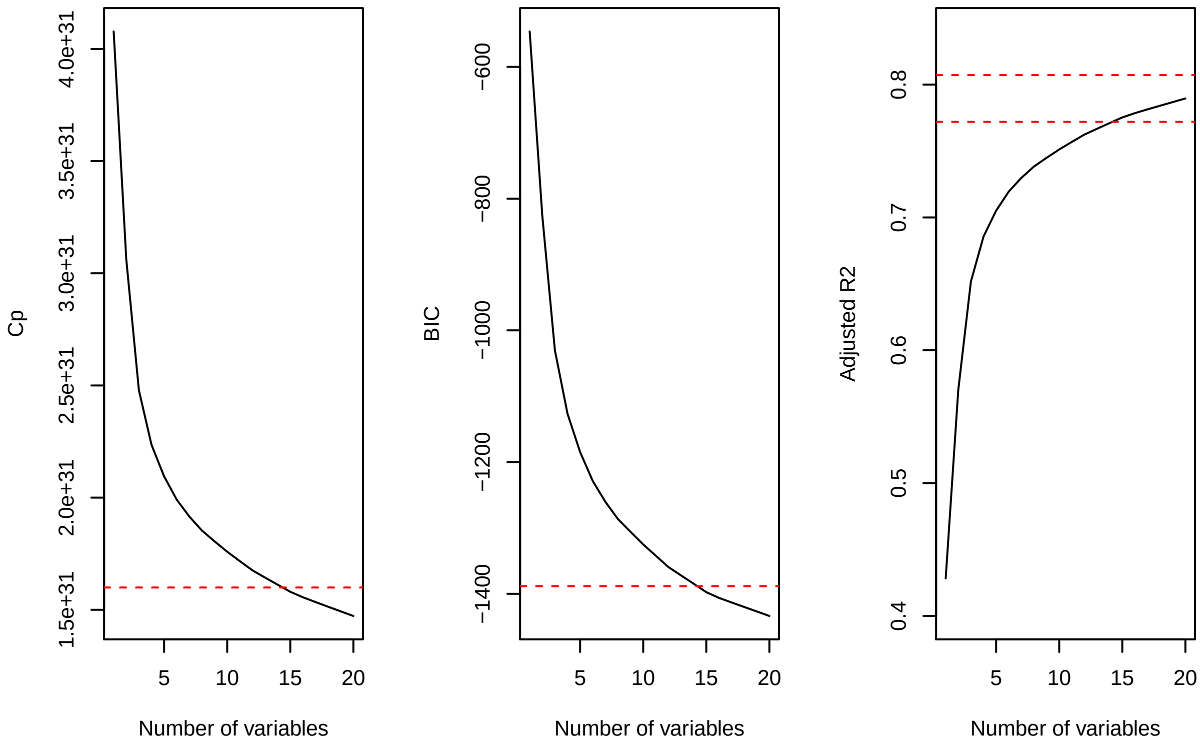
```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 160 linear dependencies found
```

```
## Reordering variables and trying again:
```

```r
fit.summary <- summary(fit)
par(mfrow = c(1, 3))
```




```r
plot(fit.summary$cp, xlab = "Number of variables", ylab = "Cp", type = "l")
min.cp <- min(fit.summary$cp)
std.cp <- sd(fit.summary$cp)
abline(h = min.cp + 0.2 * std.cp, col = "red", lty = 2)
abline(h = min.cp - 0.2 * std.cp, col = "red", lty = 2)
plot(fit.summary$bic, xlab = "Number of variables", ylab = "BIC", type='l')
min.bic <- min(fit.summary$bic)
std.bic <- sd(fit.summary$bic)
abline(h = min.bic + 0.2 * std.bic, col = "red", lty = 2)
abline(h = min.bic - 0.2 * std.bic, col = "red", lty = 2)
plot(fit.summary$adjr2, xlab = "Number of variables", ylab = "Adjusted R2", type = "l", ylim = c(0.4, 0
max.adjr2 <- max(fit.summary$adjr2)
std.adjr2 <- sd(fit.summary$adjr2)
abline(h = max.adjr2 + 0.2 * std.adjr2, col = "red", lty = 2)
abline(h = max.adjr2 - 0.2 * std.adjr2, col = "red", lty = 2)
```



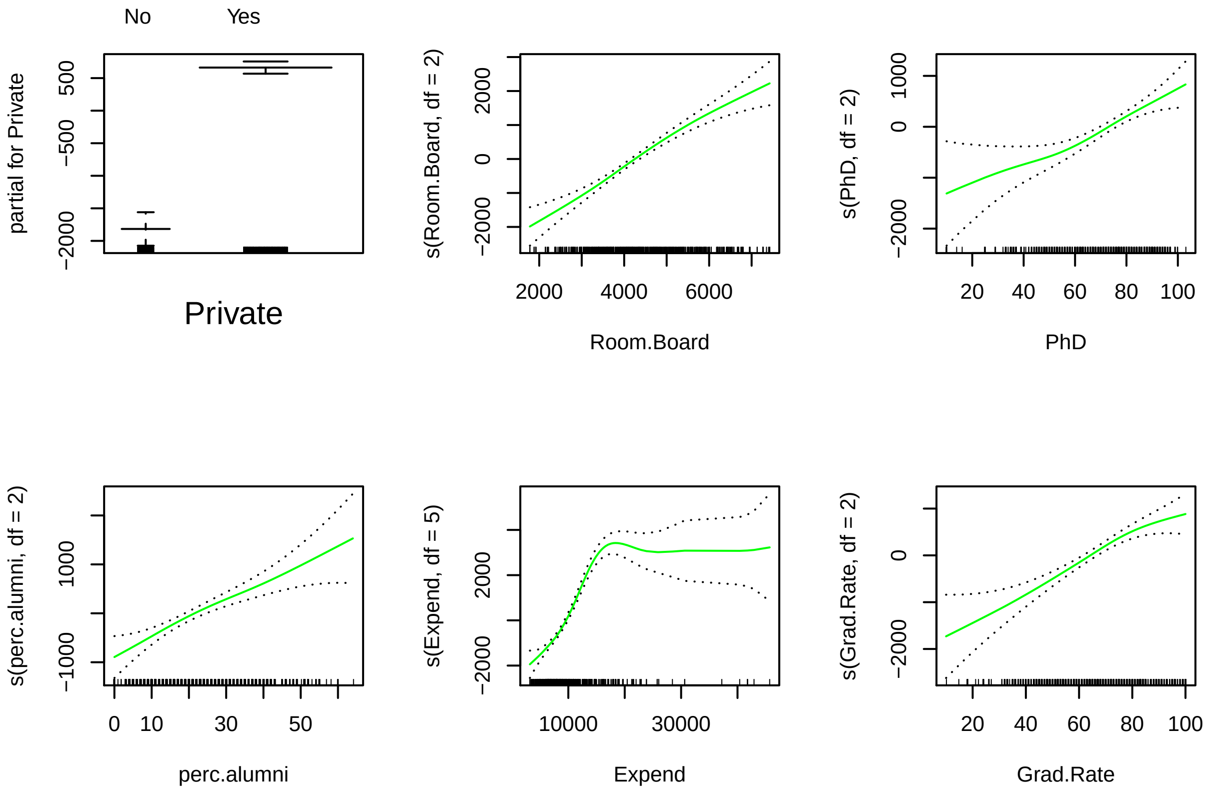

```r
# The plots of CP,BIC and Adj-R^2 show that it require minimun 14 subsets
# so that score can be within 0.2 standard deviation from the optimal.

fit <- regsubsets(Outstate ~ ., nvmax=19,data = college, method = "forward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 17 linear dependencies found
```

```r
coeffs <- coef(fit, id = 14)
names(coeffs)
```

```
##  [1] "(Intercept)"
##  [2] "School.NameBennington College"
##  [3] "School.NameBrigham Young University at Provo"
##  [4] "School.NameCreighton University"
##  [5] "School.NameGettysburg College"
##  [6] "School.NameLivingstone College"
##  [7] "School.NameUniversity of Alabama at Birmingham"
##  [8] "School.NameUniversity of Vermont"
##  [9] "School.NameWake Forest University"
## [10] "PrivateYes"
## [11] "Room.Board"
## [12] "Terminal"
## [13] "perc.alumni"
## [14] "Expend"
## [15] "Grad.Rate"
```

```r
# b)
fit <- gam(Outstate ~ Private + s(Room.Board, df = 2) + s(PhD, df = 2) +
                   s(perc.alumni, df = 2) + s(Expend, df = 5) +
                   s(Grad.Rate, df = 2), data=college.train)
par(mfrow = c(2, 3))
plot(fit, se = T, col = "green")
```




```r
# Room.Board vs s, perc.alumni vs s and grad.rate vs s look linear compare to the others.
# PhD vs s, look slightly non-linear.
# Expend vs s looks highly non-linear.

# c)
pred <- predict(fit, college.test)
(error <- mean((college.test$Outstate - pred)^2))
```

```
## [1] 3318070
```

```r
sst <- mean((college.test$Outstate - mean(college.test$Outstate))^2)
rss <- 1 - error / sst
rss
```

```
## [1] 0.7845758
```

```r
#  GAM with 14 predictors we obtained test R-squared is 0.785. This result is has
# a little improvement towards OLS.
```

```r
# d)
summary(fit)
```

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 2) + s(PhD,
##     df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate,
##     df = 2), data = college.train)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -7430.9 -1214.1   138.8  1324.7  8367.3
##
## (Dispersion Parameter for gaussian family taken to be 3800359)
##
##     Null Deviance: 15321315691 on 999 degrees of freedom
## Residual Deviance: 3743355264 on 985.0006 degrees of freedom
## AIC: 18005.37
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                         Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Private                  1 4139572514 4139572514 1089.258 < 2.2e-16 ***
## s(Room.Board, df = 2)    1 3178728668 3178728668  836.429 < 2.2e-16 ***
## s(PhD, df = 2)           1 1026089713 1026089713  269.998 < 2.2e-16 ***
## s(perc.alumni, df = 2)   1  587330769  587330769  154.546 < 2.2e-16 ***
## s(Expend, df = 5)        1  983457788  983457788  258.780 < 2.2e-16 ***
## s(Grad.Rate, df = 2)     1  176292574  176292574   46.388 1.685e-11 ***
## Residuals              985 3743355264    3800359
## ---
## 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 = 2)        1  3.555 0.05966 .
## s(PhD, df = 2)               1  2.840 0.09229 .
## s(perc.alumni, df = 2)       1  0.975 0.32374
## s(Expend, df = 5)            4 33.829 < 2e-16 ***
## s(Grad.Rate, df = 2)         1  4.658 0.03116 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Non- parametric ANOVA approach shows that there is strong non-linear relationship between
# response variable and the predictor expend.
```