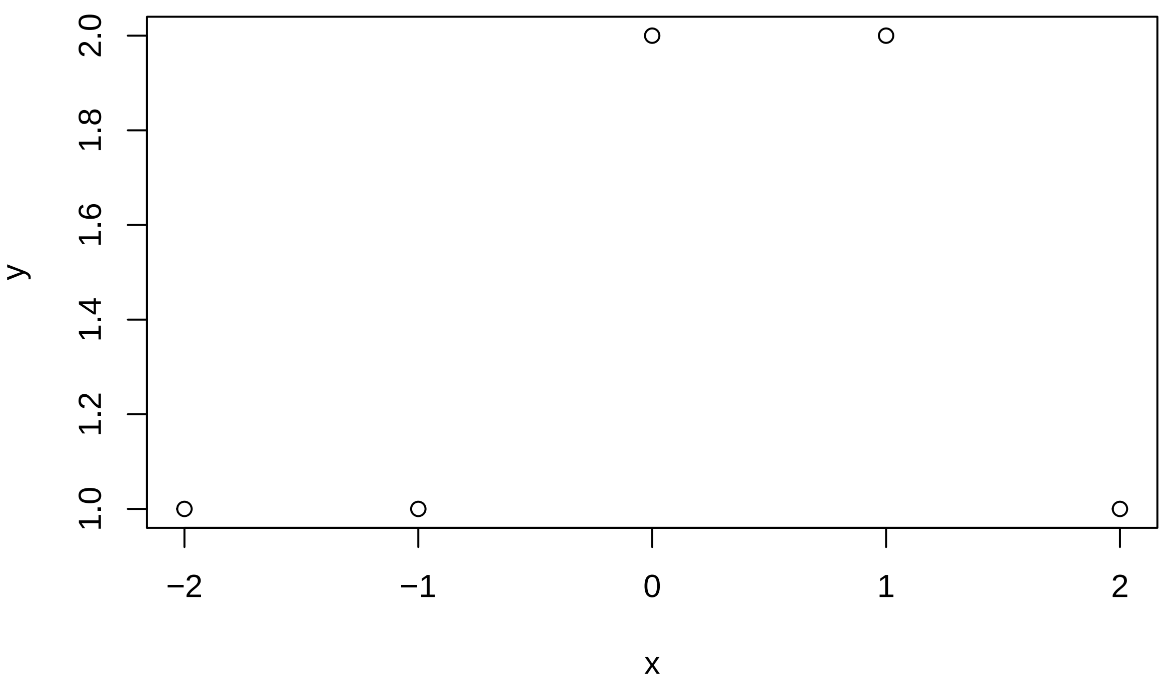
101c HW 5

Lucy Shao, 304575686

May 24, 2017

Textbook Problems: ###1. Question 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)
```



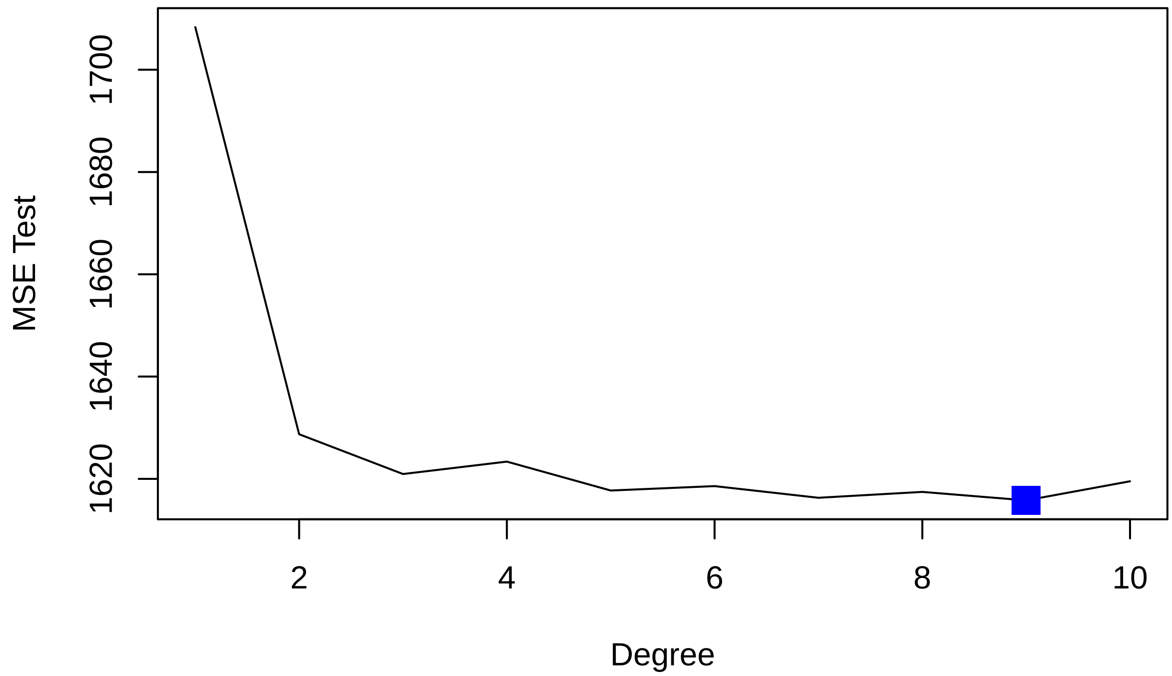
The curve have 3 parts, function between -2 and 0 is y=1, the function between 0 and 1 is y =2, the function between 1 and 2 is y=3-x.

2. Question 6

```
set.seed(123)
#(a)
wagedata<-read.csv("/Users/lucy/Downloads/WageLec2.csv")
attach(wagedata)
library(boot)
```

```
## Warning: package 'boot' was built under R version 3.3.2
```

```r
library(ISLR)
deltas<-rep(NA,10)
for (i in 1:10){
fit<-glm(wage~poly(age,i),data=wagedata)
deltas[i]<-cv.glm(wagedata,fit,K=10)$delta[1] }
plot(1:10,deltas,xlab="Degree",ylab="MSE Test",type="l")
min1<-which.min(deltas)
points(min1,deltas[min1],col="blue",cex=2,pch=15)
```

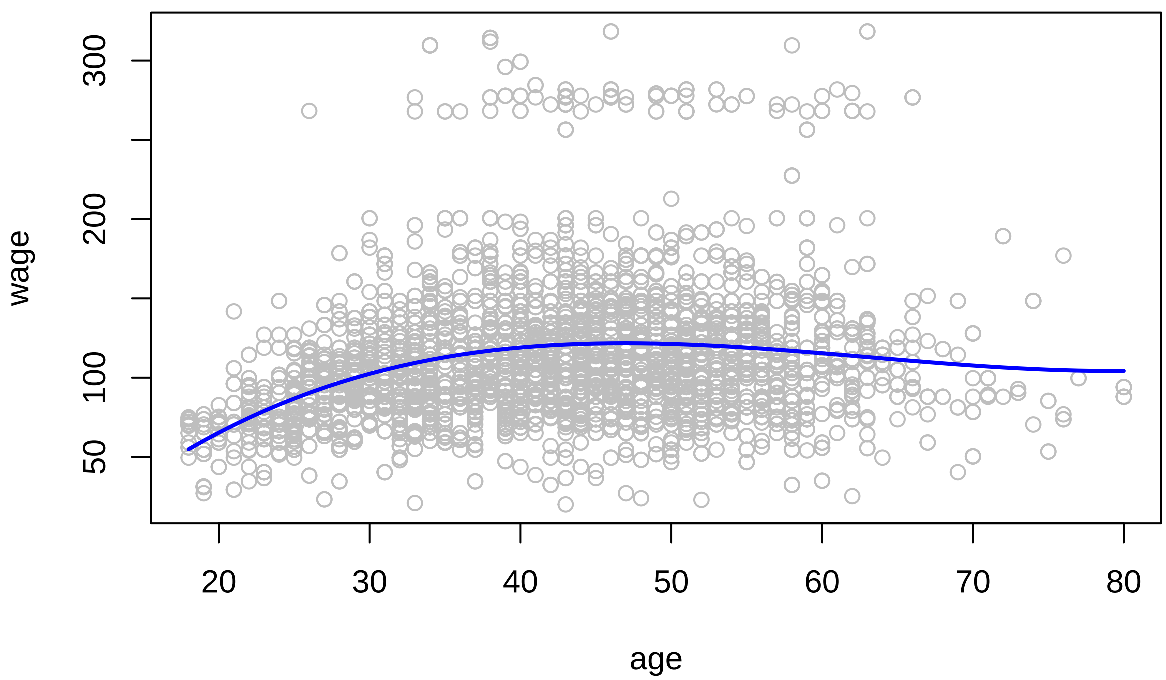



```r
# degree 9 is the optimal
fit1 <- lm(wage ~ age, data = wagedata)
fit2 <- lm(wage ~ poly(age, 2), data = wagedata)
fit3 <- lm(wage ~ poly(age, 3), data = wagedata)
fit4 <- lm(wage ~ poly(age, 4), data = wagedata)
fit5 <- lm(wage ~ poly(age, 5), data = wagedata)
fit6 <- lm(wage ~ poly(age, 6), data = wagedata)
fit7 <- lm(wage ~ poly(age, 7), data = wagedata)
fit8 <- lm(wage ~ poly(age, 8), data = wagedata)
fit9 <- lm(wage ~ poly(age, 9), data = wagedata)
anova(fit1, fit2, fit3, fit4, fit5,fit6,fit7,fit8,fit9)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## 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)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   3998 6827523
## 2   3997 6503669  1    323854 200.6162 < 2.2e-16 ***
## 3   3996 6471970  1     31699  19.6361  9.62e-06 ***
## 4   3995 6469894  1      2076   1.2859  0.256871
## 5   3994 6457099  1     12795   7.9260  0.004897 **
## 6   3993 6452761  1      4339   2.6876  0.101213
## 7   3992 6446093  1      6668   4.1306  0.042181 *
## 8   3991 6446068  1        24   0.0151  0.902148
## 9   3990 6441046  1      5022   3.1111  0.077839 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

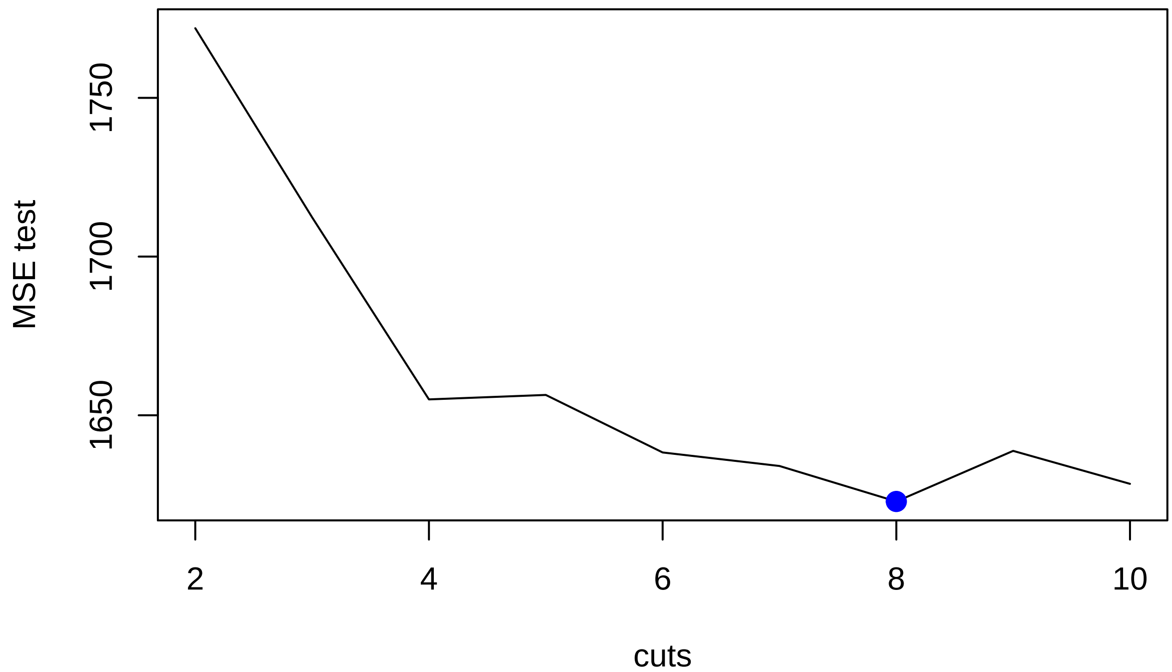
```r
#by looking at the output, degree 2 or 3 provides the most reasonable fit
plot(wage ~ age, data = wagedata, col = "grey")
agelim<- range(age)
age.grid<-seq(from=agelim[1], to = agelim[2])
fit <-lm(wage ~ poly(age, 3), data = wagedata)
pred<-predict(fit, newdata = list(age =age.grid))
lines(age.grid, pred, col = "blue", lwd = 2)
```



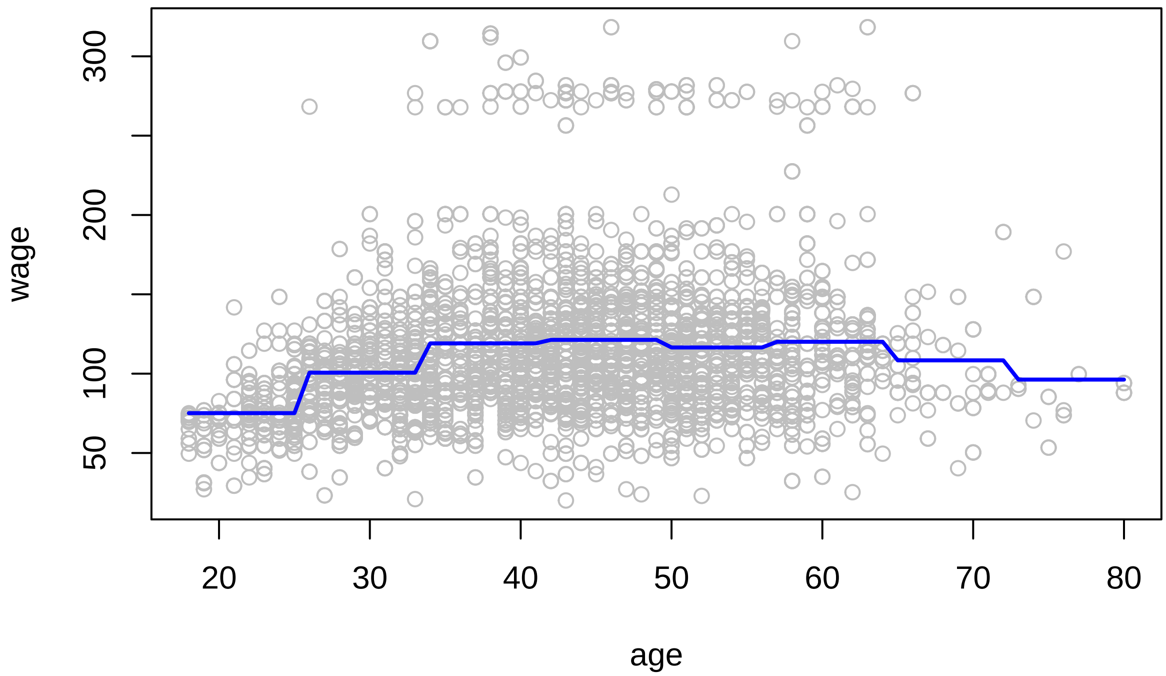
b)

```r
#(b)
cvs <-rep(NA,10)
for (i in 2:10) {
wagedata$age.cut<-cut(wagedata$age, i)
fit <- glm(wage~age.cut, data = wagedata)
cvs[i] <- cv.glm(wagedata, fit, K=10)$delta[1]
}
```

```r
plot(2:10, cvs[-1], xlab = "cuts", ylab = "MSE test", type = "l")
d.min <-which.min(cvs)
points(d.min, cvs[d.min], col = "blue", cex = 2, pch = 20)
```




```r
#Error is minimum  at 8 cuts.
plot(wage~age, data = wagedata, col = "grey")
agelim<- range(wagedata$age)
age.grid <-seq(from =agelim[1], to =agelim[2])
fit <-glm(wage ~ cut(age, 8), data = wagedata)
pred <- predict(fit, data.frame(age = age.grid))
lines(age.grid, pred, col = "blue", lwd = 2)
```



3. Question 7

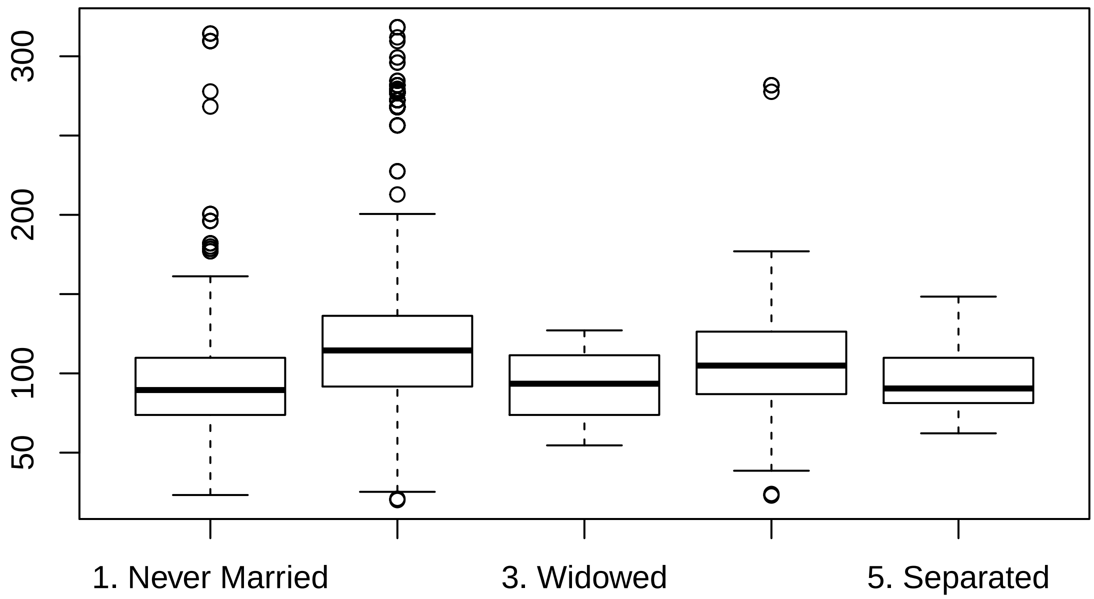
```
set.seed(123)
summary(wagedata$maritl)
```

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

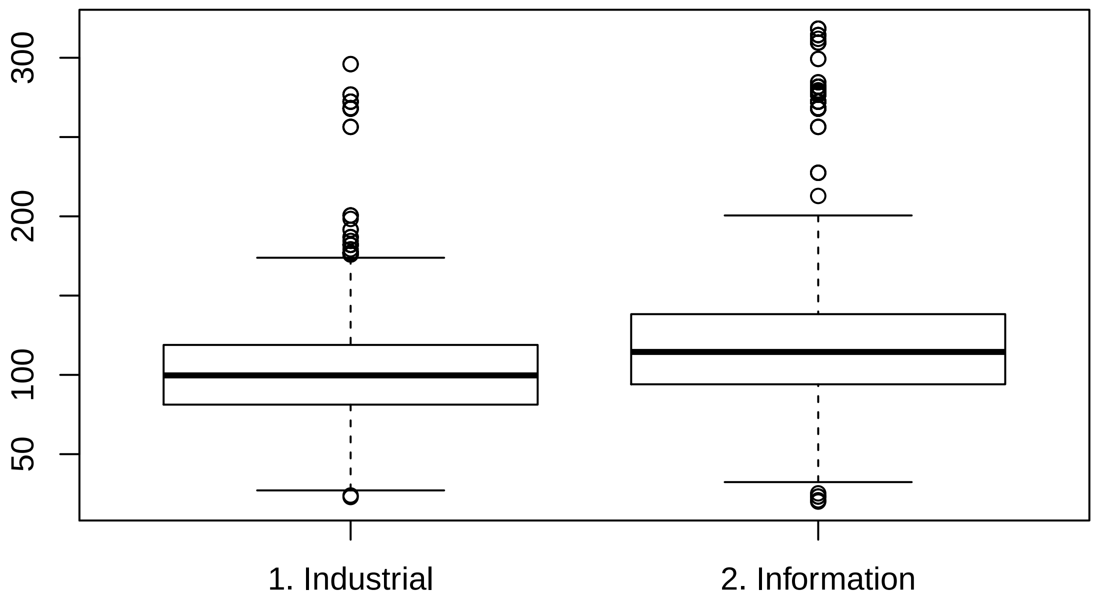
```
summary(wagedata$jobclass)
```

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

```
plot(wagedata$maritl,wagedata$wage)
```




```r
plot(wagedata$jobclass,wagedata$wage)
```




```r
#So in general, married couples earn more on average, informational jobs earn more on average
library(gam)
```

```
## Warning: package 'gam' was built under R version 3.3.2
```

```
## 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 = wagedata)
deviance(fit0)
```

```
## [1] 4942702
```

```r
fit1 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education+jobclass, data = wagedata)
deviance(fit1)
```

```
## [1] 4909333
```

```r
fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education+maritl, data = wagedata)
deviance(fit2)
```

```
## [1] 4845935
```

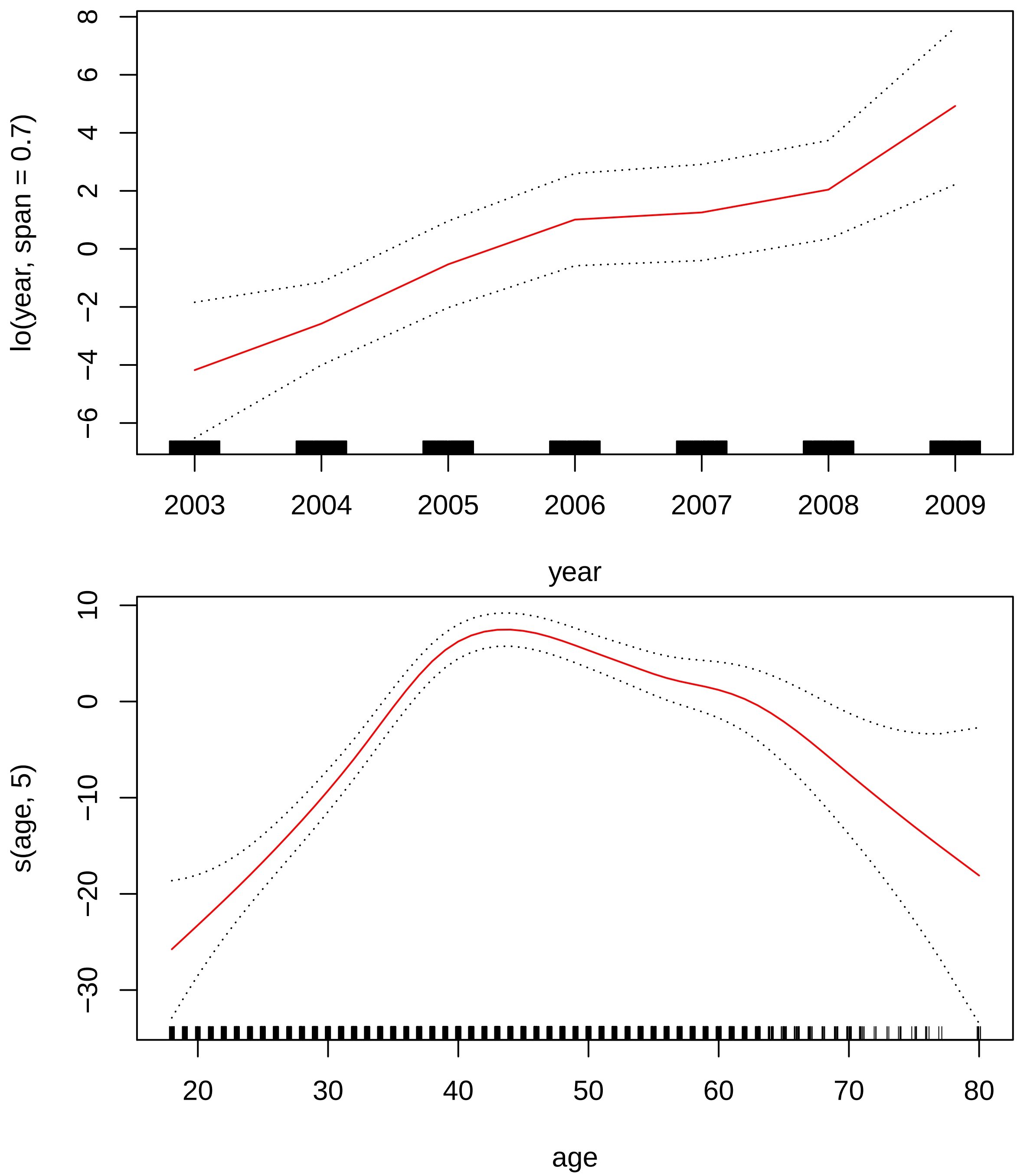
```r
fit3 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education+ jobclass + maritl, data = wagedata)
deviance(fit3)
```

```
## [1] 4807302
```

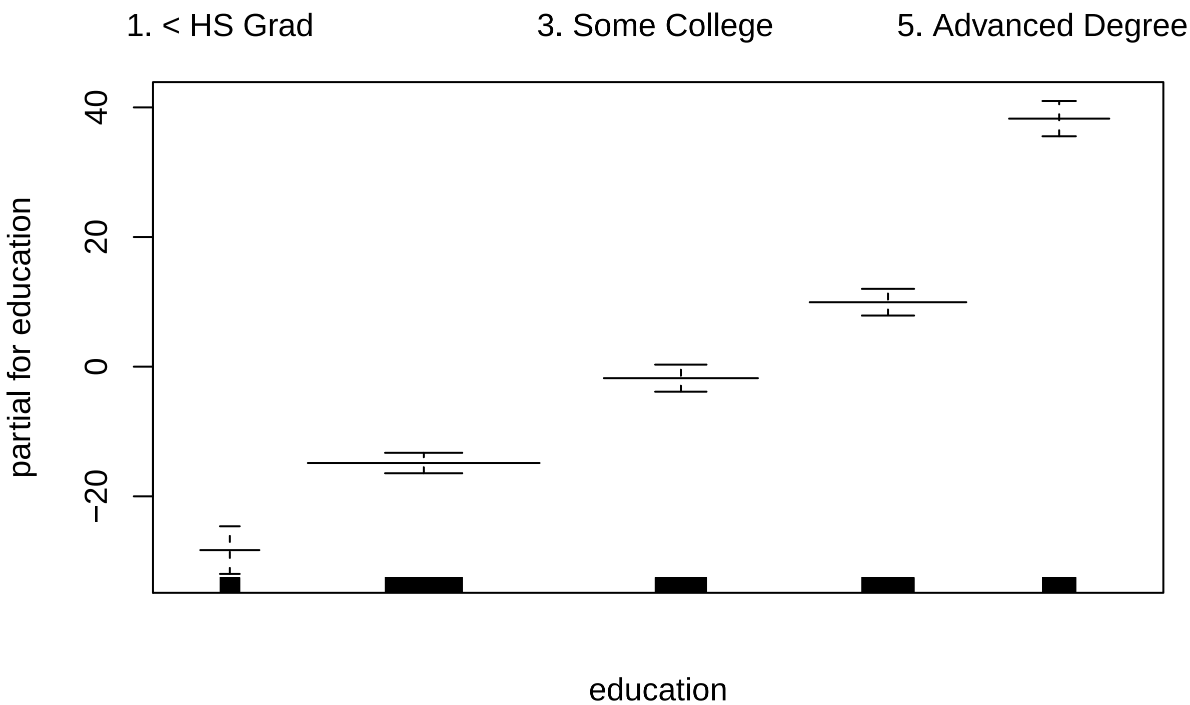
```r
anova(fit0, fit1, fit2, fit3)
```

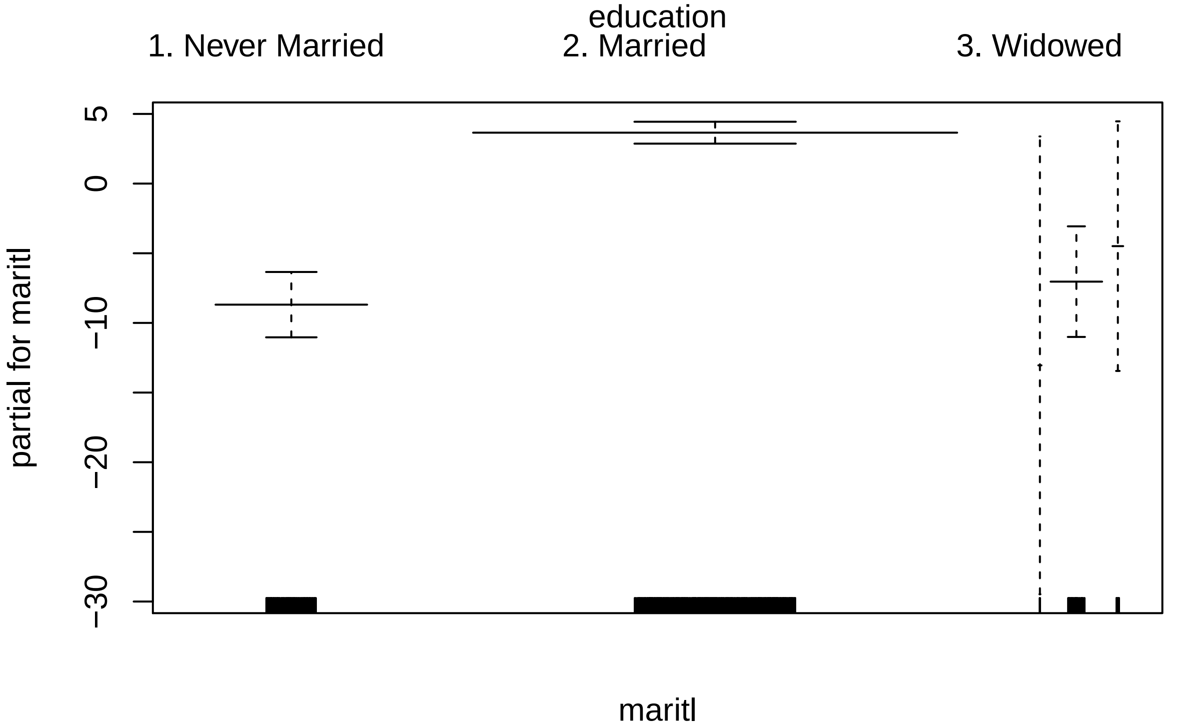
```
## Analysis of Deviance Table
##
## Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
## Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
## Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
## Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + 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
#So we choose fit2 by looking at the anova
plot(fit2,se=TRUE,col="red")
```



lo(year, span = 0.7)
8
6
4
2
0
−2
−4
−6
2003
2004
2005
2006
2007
2008
2009
year
s(age, 5)
10
0
−10
−20
−30
20
30
40
50
60
70
80
age





4. Question 10

```
College<-read.csv("/Users/lucy/Downloads/CollegeLec2.csv")
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```r
set.seed(123)
attach(College)
```

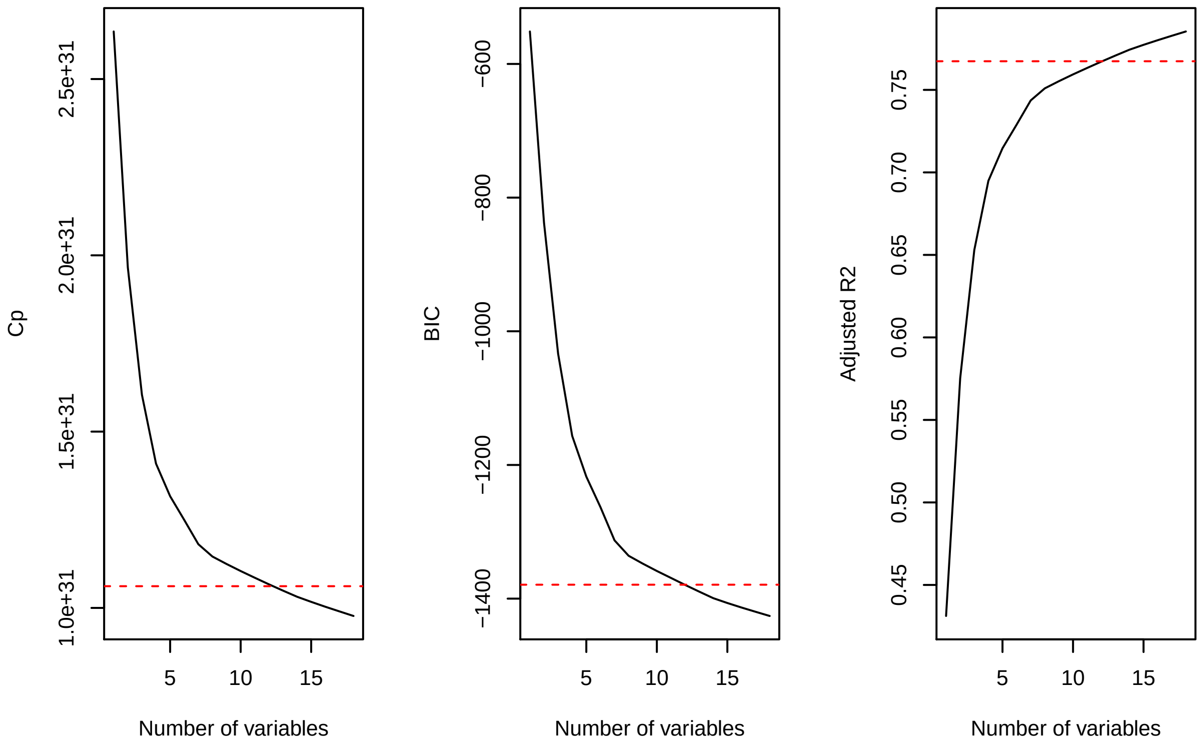
```
## The following object is masked from wagedata:
##
##     X
```

```r
train <- sample(length(Outstate), length(Outstate) / 2)
test <- -train
College.train <- College[train, ]
College.test <- College[test, ]
fit <- regsubsets(Outstate ~ ., data = College.train, nvmax = 17, method = "forward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 174 linear dependencies found
```

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

```r
fit.summary <- summary(fit)
par(mfrow = c(1, 3))
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$adjr2, xlab = "Number of variables", ylab = "Adjusted R2", type = "l")
max.adjr2 <- max(fit.summary$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)
```



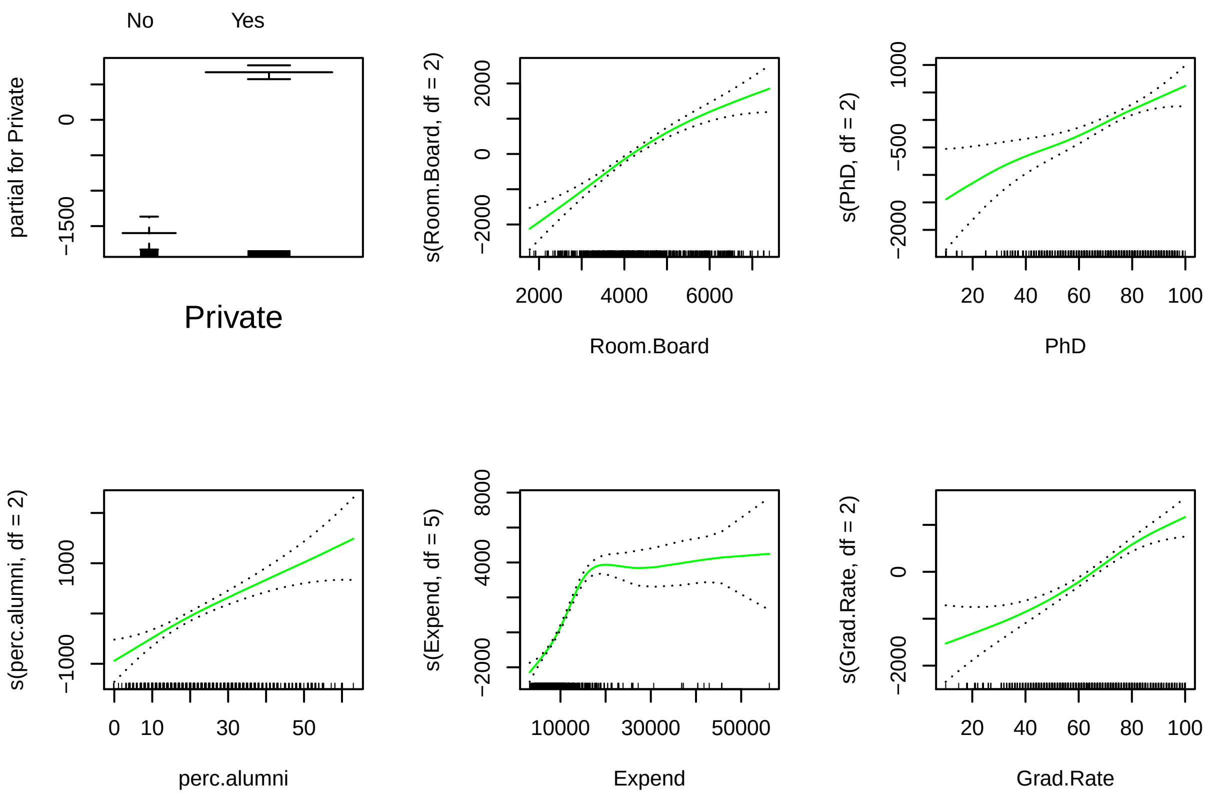

```r
fit <- regsubsets(Outstate ~ ., data = College, method = "forward")
```

```
## 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=6)
names(coeffs)
```

```
## [1] "(Intercept)" "PrivateYes"  "Room.Board"  "Terminal"    "perc.alumni"
## [6] "Expend"      "Grad.Rate"
```

```r
#b
#
library(gam)
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
#c
#
preds <- predict(fit, College.test)
err <- mean((College.test$Outstate - preds)^2)
err
```

```
## [1] 3640187
```

```r
tss <- mean((College.test$Outstate - mean(College.test$Outstate))^2)
rss <- 1 - err / tss
rss
```

```
## [1] 0.7628301
```

```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
## -7769.72 -1148.73    93.57  1167.75  8020.91
##
## (Dispersion Parameter for gaussian family taken to be 3510922)
```

```
##
##     Null Deviance: 15372053414 on 999 degrees of freedom
## Residual Deviance: 3458259860 on 985.0005 degrees of freedom
## AIC: 17926.15
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                         Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Private                  1 3900475941 3900475941 1110.955 < 2.2e-16 ***
## s(Room.Board, df = 2)    1 3094953438 3094953438  881.521 < 2.2e-16 ***
## s(PhD, df = 2)           1 1025925271 1025925271  292.210 < 2.2e-16 ***
## s(perc.alumni, df = 2)   1  599258115  599258115  170.684 < 2.2e-16 ***
## s(Expend, df = 5)        1 1192562100 1192562100  339.672 < 2.2e-16 ***
## s(Grad.Rate, df = 2)     1  224677193  224677193   63.994 3.488e-15 ***
## Residuals              985 3458259860    3510922
## ---
## 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  6.541 0.01069 *
## s(PhD, df = 2)               1  2.022 0.15536
## s(perc.alumni, df = 2)       1  0.657 0.41765
## s(Expend, df = 5)            4 44.303 < 2e-16 ***
## s(Grad.Rate, df = 2)         1  5.428 0.02002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Non-parametric Anova test shows a strong evidence of non-linear relationship between response and Expend, and a moderately strong non-linear relationship (using p value of 0.05) between response and Grad.Rate or PhD.