Exercise VI.7.9.7

Kaarthik Sundaramoorthy, Sahil Shah and Vidhi Shah

7/4/2020

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.

Importing the Dataset

```
library(ISLR)
data(Wage)
attach(Wage)
```

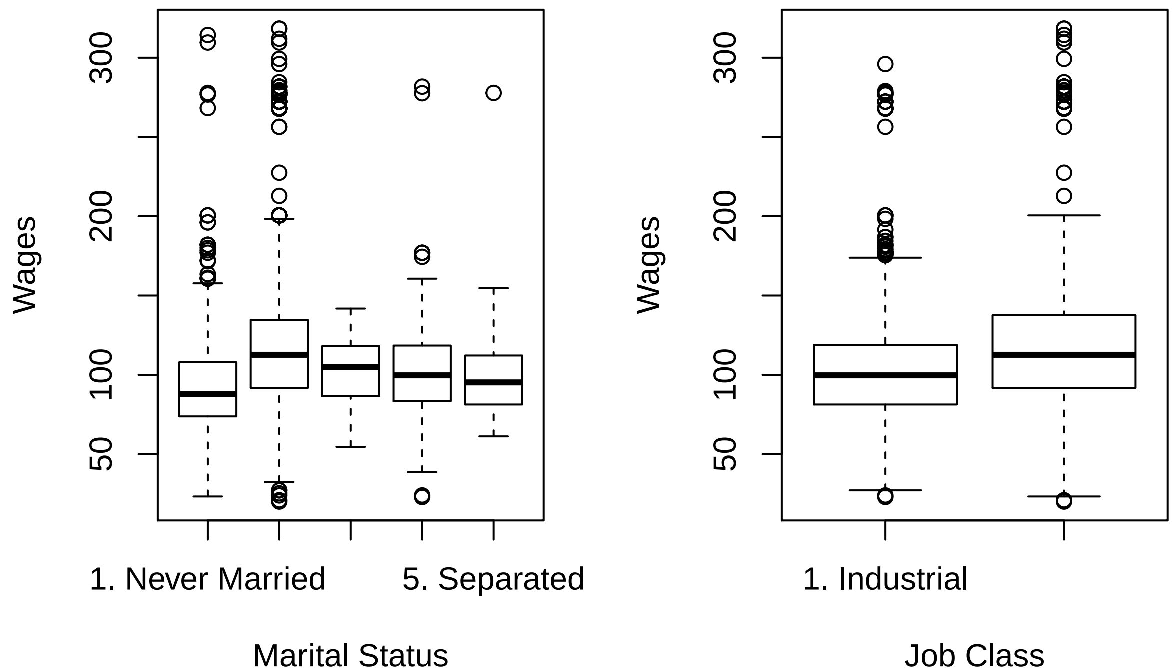
Exploring the features **maritl**, **jobclass**, **health**, and **health_ins** with the help of `summary()`

```
set.seed(123)
summary(Wage[, c("maritl", "jobclass", "health", "health_ins")] )
```

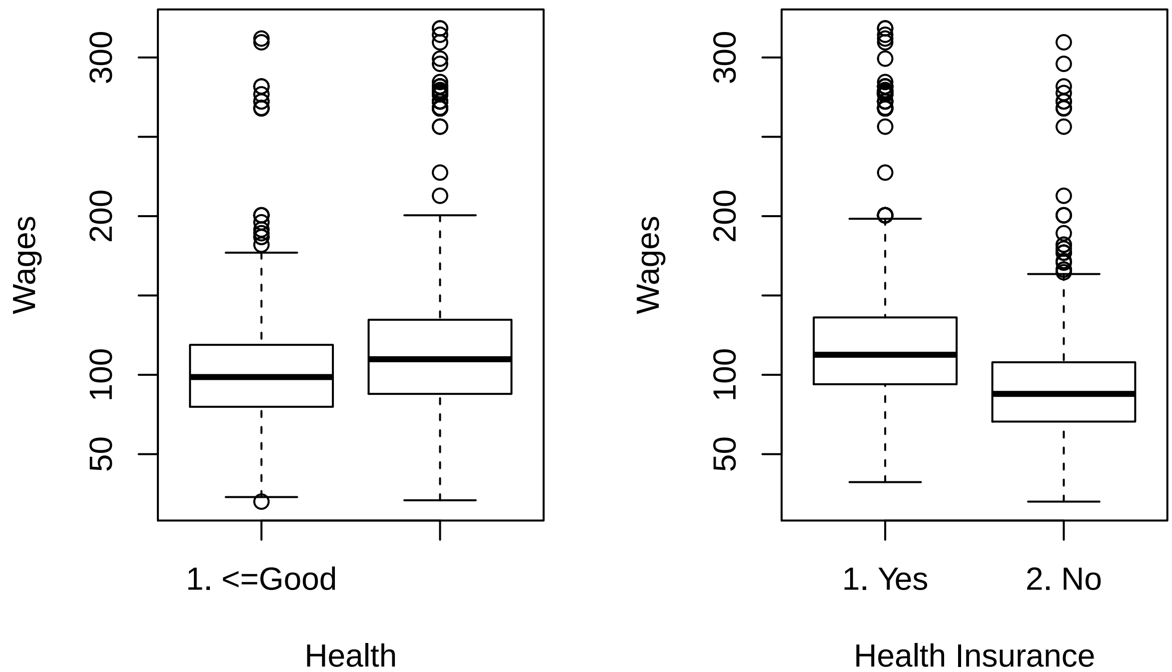
```
##               maritl               jobclass                health
##  1. Never Married: 648   1. Industrial :1544   1. <=Good     : 858
##  2. Married      :2074   2. Information:1456   2. >=Very Good:2142
##  3. Widowed      :  19
##  4. Divorced     : 204
##  5. Separated    :  55
##   health_ins
##  1. Yes:2083
##  2. No : 917
##
##
##
```

Generating Box-Plots

```
#attach(Wage)
par(mfrow = c(1, 2))
plot(maritl, wage, xlab = "Marital Status", ylab = "Wages")
plot(jobclass, wage, xlab = "Job Class", ylab = "Wages")
```




```r
plot(health, wage, xlab = "Health", ylab = "Wages")
plot(health_ins, wage, xlab = "Health Insurance", ylab = "Wages")
```



maritl, **jobclass**, **health** and **health_ins** are categorical variables. From the plots it can be said that the couple who are married earns more than other three categories and Informational wage earns more on average.From the plots we also see that better the wages, better is the health. Health insurance is directly related to wages. People having health insurance have more wages.

Generating the polynomial and step function for the same.

```r
poly_fit = lm(wage ~ maritl, data = Wage)
#deviance(poly_fit)
poly_fit1 = lm(wage ~ jobclass, data = Wage)
#deviance(poly_fit1)
poly_fit2 = lm(wage ~ maritl + jobclass, data = Wage)
#deviance(poly_fit2)
anova(poly_fit, poly_fit1, poly_fit2)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ maritl
## Model 2: wage ~ jobclass
## Model 3: wage ~ maritl + jobclass
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2995 4858941
## 2   2998 4998547 -3   -139606 29.932 < 2.2e-16 ***
## 3   2994 4654752  4    343795 55.283 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The sample data fit is minimized with the most complex linear model.

We will use `gam()` to predict the **wage** using the splines of the **year**, **education**, **age**, **maritl**, and **jobclass**. Because we can't fit splines in the categorical values.

```
library(gam)
```

```
## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.16.1
```

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

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
fit2 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass, data = Wage)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

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

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
fit4 <- gam(wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass + maritl, data = Wage)
```

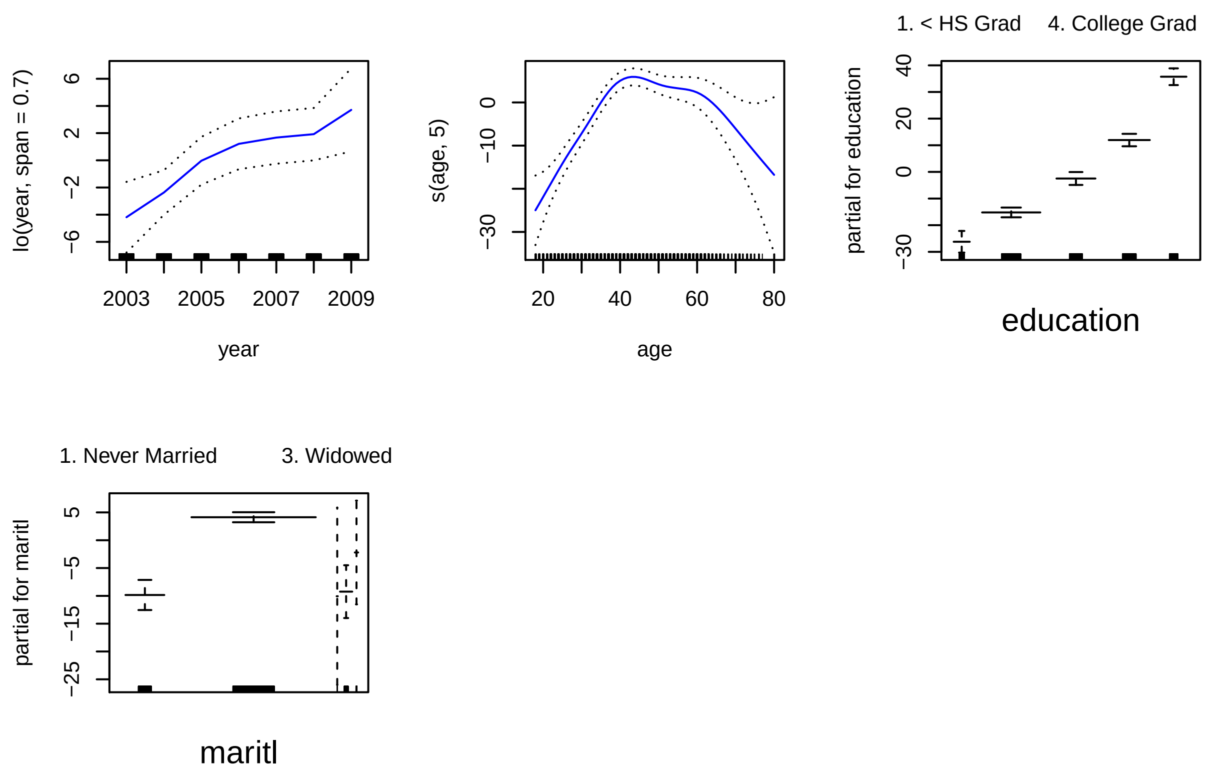
```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
anova(fit1, fit2, fit3, fit4)
```

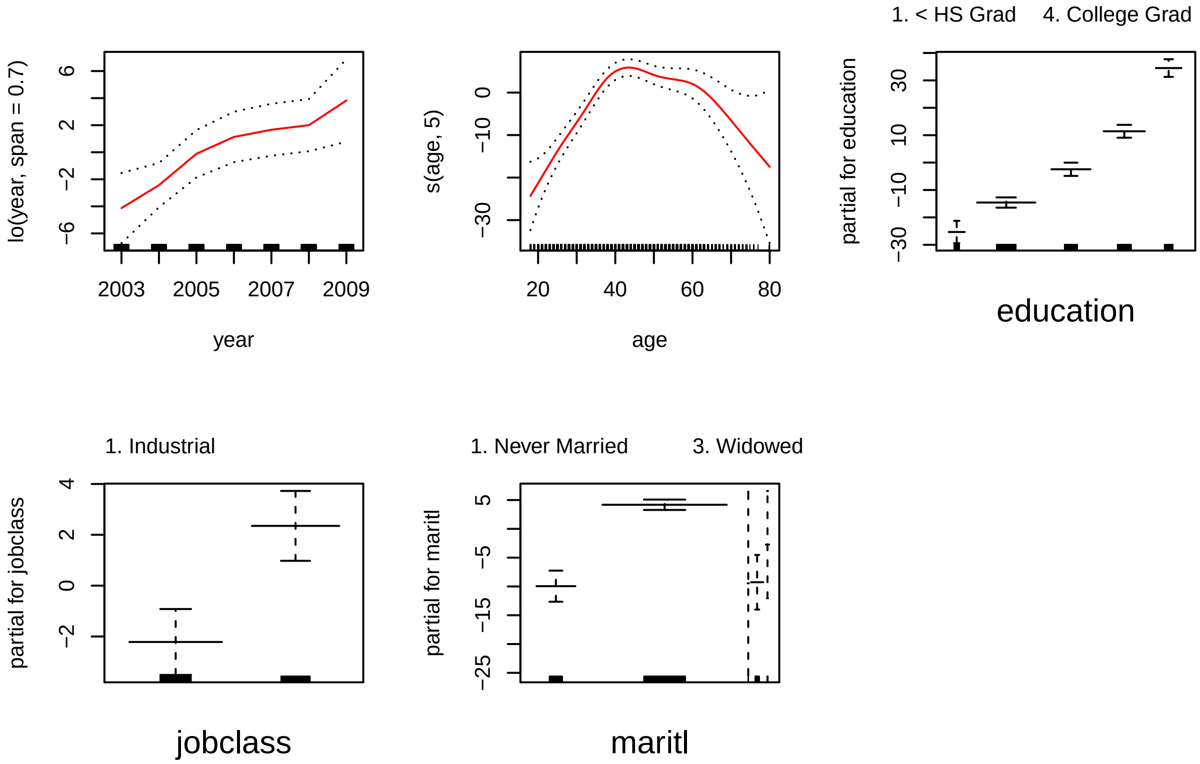
```
## Analysis of Deviance Table
##
## Model 1: wage ~ lo(year, span = 0.7) + s(age, 5) + education
## Model 2: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass
## Model 3: wage ~ lo(year, span = 0.7) + s(age, 5) + education + maritl
## Model 4: wage ~ lo(year, span = 0.7) + s(age, 5) + education + jobclass +
##     maritl
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1    2987.1    3691855
## 2    2986.1    3679689  1    12166 0.0014637 **
## 3    2983.1    3597526  3    82163  9.53e-15 ***
## 4    2982.1    3583675  1    13852 0.0006862 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is obvious that **age** has non-linear relationship with **wage**. Besides, there are acceptable information that **wage** feature differs from the given different qualitative predictors including **education**, **jobclass**, **maritl**. We also discovered that **year** has an interaction effect with **age** therefore we don't have to add individual basis functions for **year** and **age**. According to the results the model 3 and model 4 can be considered as the significantly better.

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




```
plot(fit4, se = T, col = "red")
```


lo(year, span = 0.7)
6
2
−2
−6
2003
2005
2007
2009
year
s(age, 5)
0
−10
−30
20
40
60
80
age
1. < HS Grad
4. College Grad
partial for education
30
10
−10
−30
education
1. Industrial
partial for jobclass
4
2
0
−2
jobclass
1. Never Married
3. Widowed
partial for maritl
5
−5
−15
−25
maritl