Assignment 4

79546 - Stephen K. Ng'etich

Contents

1 Pre-requisite

1.1 Load Packages

```
# Clear variables
rm(list=ls())

library(splines)
library(npreg)
library(ISLR)
library(dplyr)
library(ggplot2)
library(gam)
```

1.2 Load Dataset

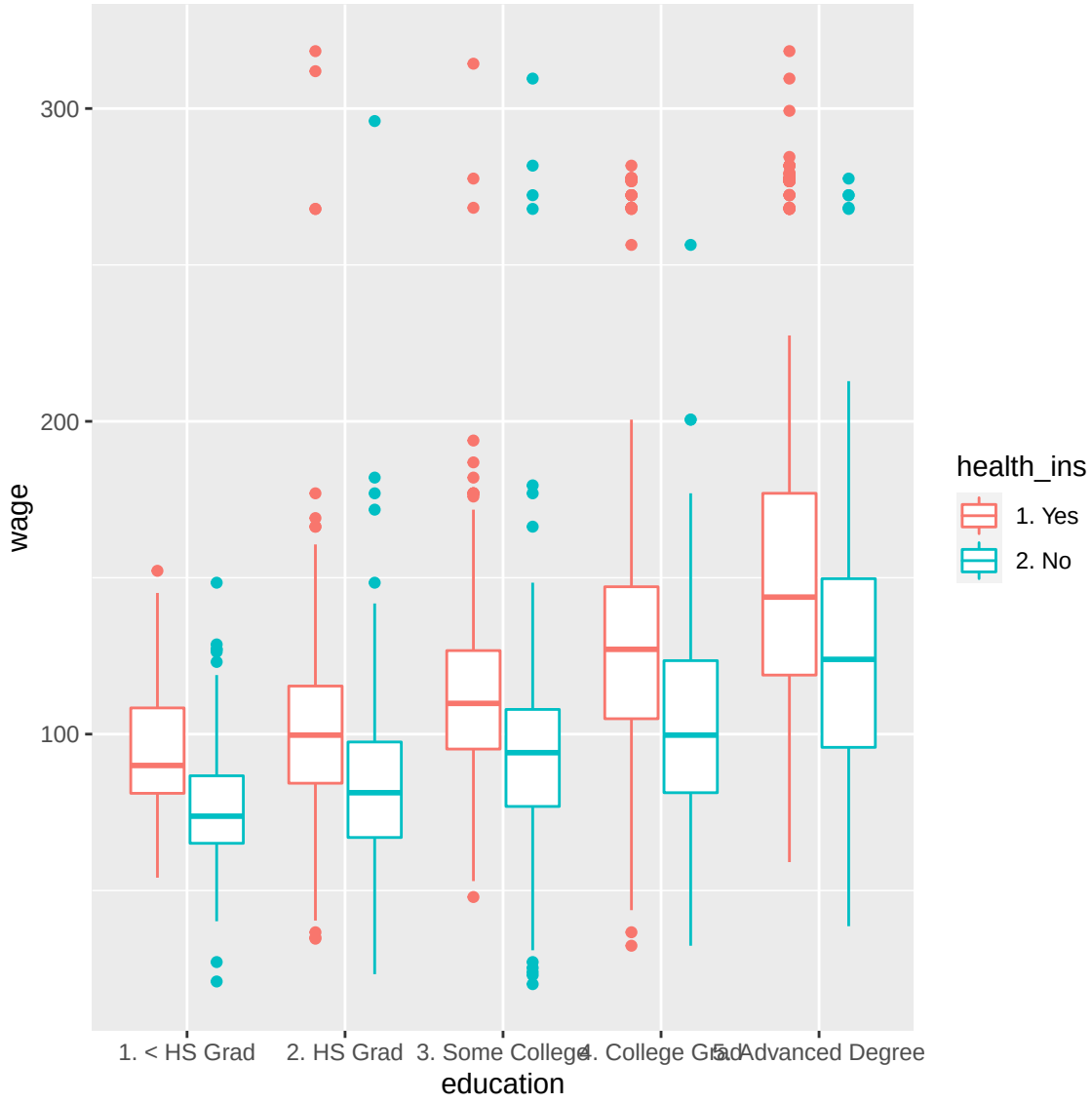
```
summary(Wage)
```

```
##       year           age                     maritl           race
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##               education                     region               jobclass
##  1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544
##  2. HS Grad        :971   1. New England       :   0   2. Information:1456
##  3. Some College   :650   3. East North Central:   0
##  4. College Grad   :685   4. West North Central:   0
##  5. Advanced Degree:426   5. South Atlantic    :   0
##                           6. East South Central:   0
##                           (Other)              :   0
##             health      health_ins      logwage           wage
##  1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
##  2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
##                                      Median :4.653   Median :104.92
##                                      Mean   :4.654   Mean   :111.70
##                                      3rd Qu.:4.857   3rd Qu.:128.68
##                                      Max.   :5.763   Max.   :318.34
##
```

```
dataset = Wage
```

2 Exploratory Data Analysis

```
ggplot(Wage, aes(x = education, y = wage, color = health_ins)) +  # ggplot function
  geom_boxplot()
```



3 Model Selection

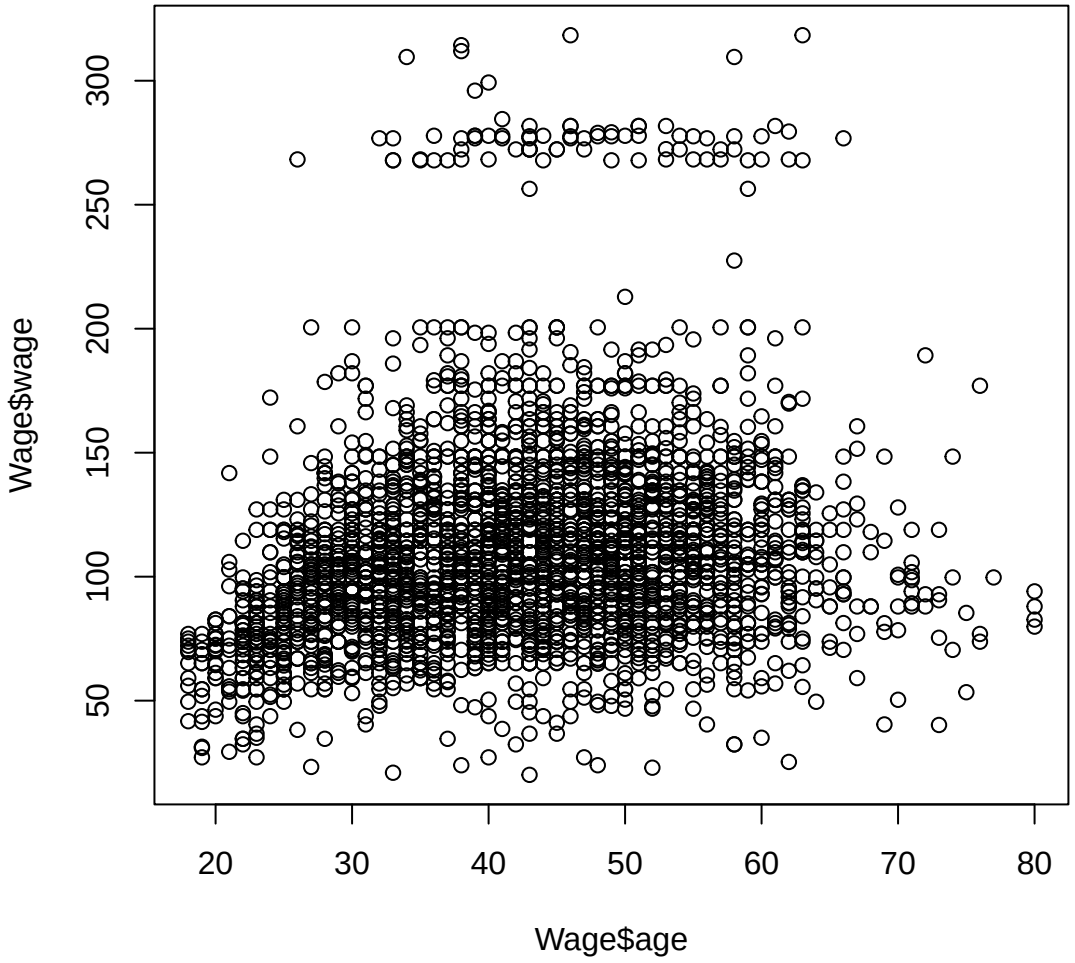
```
lm_model =  lm(wage ~ year+age ,data = dataset)
summary(lm_model)
```

```
##
## Call:
## lm(formula = wage ~ year + age, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.766 -25.081  -6.108  16.838 209.053
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2318.5309   739.1385  -3.137  0.00172 **
## year            1.1968     0.3685   3.247  0.00118 **
## age             0.6992     0.0647  10.808  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 2997 degrees of freedom
## Multiple R-squared:  0.04165,    Adjusted R-squared:  0.04101
## F-statistic: 65.12 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Generate a plot of the wages over age

```
plot(Wage$age,Wage$wage)
```



Get min/max values of age using the range() function

```
agelims = Wage %>%
    select(age) %>%
    range

agelims
```

```
## [1] 18 80
```

Generate a sequence of age values spanning the range

```
# Generate a sequence of age values spanning the range
age_grid = seq(from = min(agelims), to = max(agelims))
age_grid
```

```
##  [1] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
## [26] 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67
## [51] 68 69 70 71 72 73 74 75 76 77 78 79 80
```

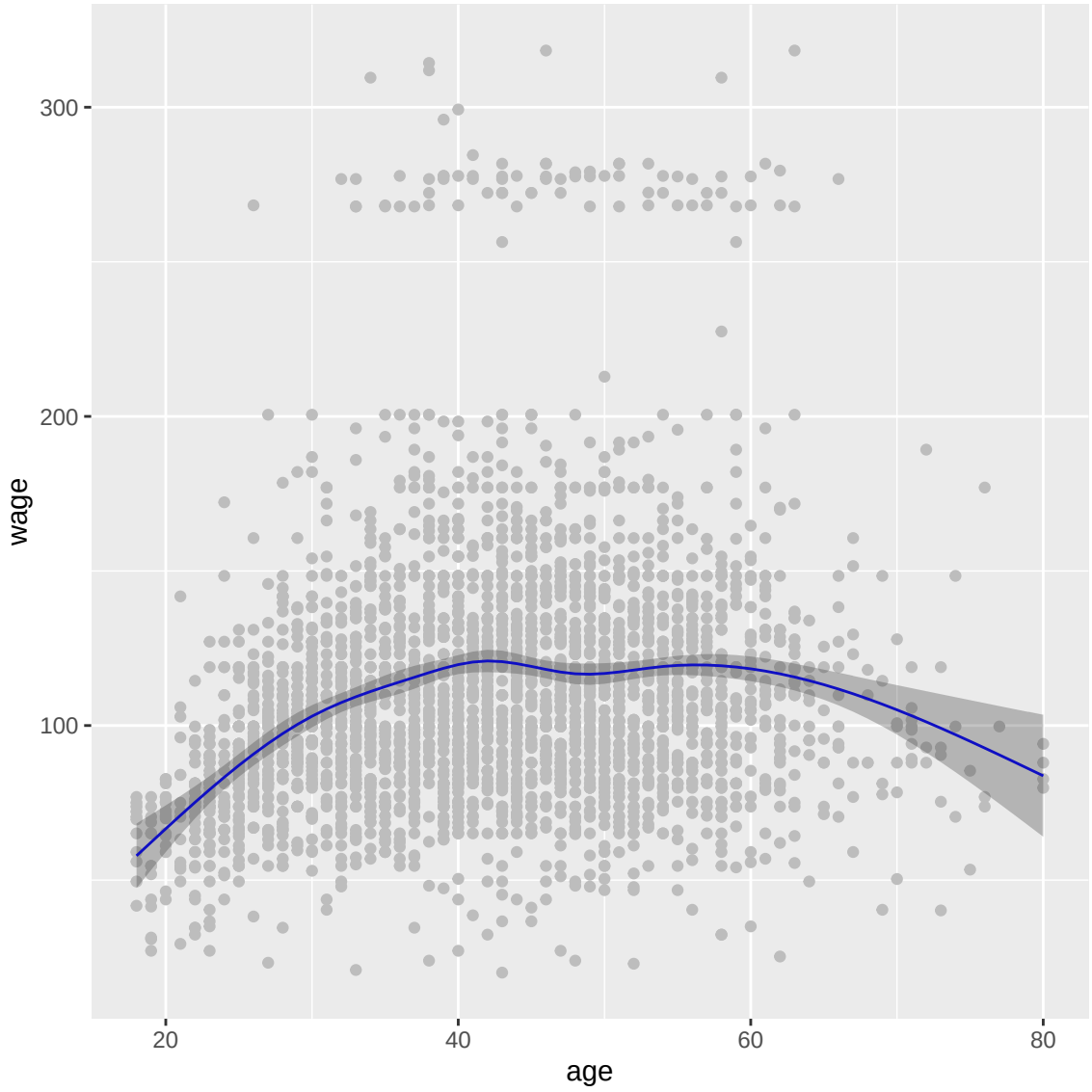
3.1 Regression Splines

```
# Fit a regression spline using basis functions
fit =  lm(wage~ns(age,df=6), data=Wage)

# Predict the value of the generated ages,
# returning the standard error using se = TRUE
pred = predict(fit, newdata = list(age = age_grid), se = TRUE)

# Compute error bands (2*SE)
se_bands = with(pred, cbind("upper" = fit+2*se.fit,
                            "lower" = fit-2*se.fit))

# Plot the spline and error bands
ggplot() +
  geom_point(data = Wage, aes(x = age, y = wage),colour="gray74") +
  geom_line(aes(x = age_grid, y = pred$fit), color = "#0000FF") +
  geom_ribbon(aes(x = age_grid,
                  ymin = se_bands[,"lower"],
                  ymax = se_bands[,"upper"]),
              alpha = 0.3) +
  xlim(agelims)
```

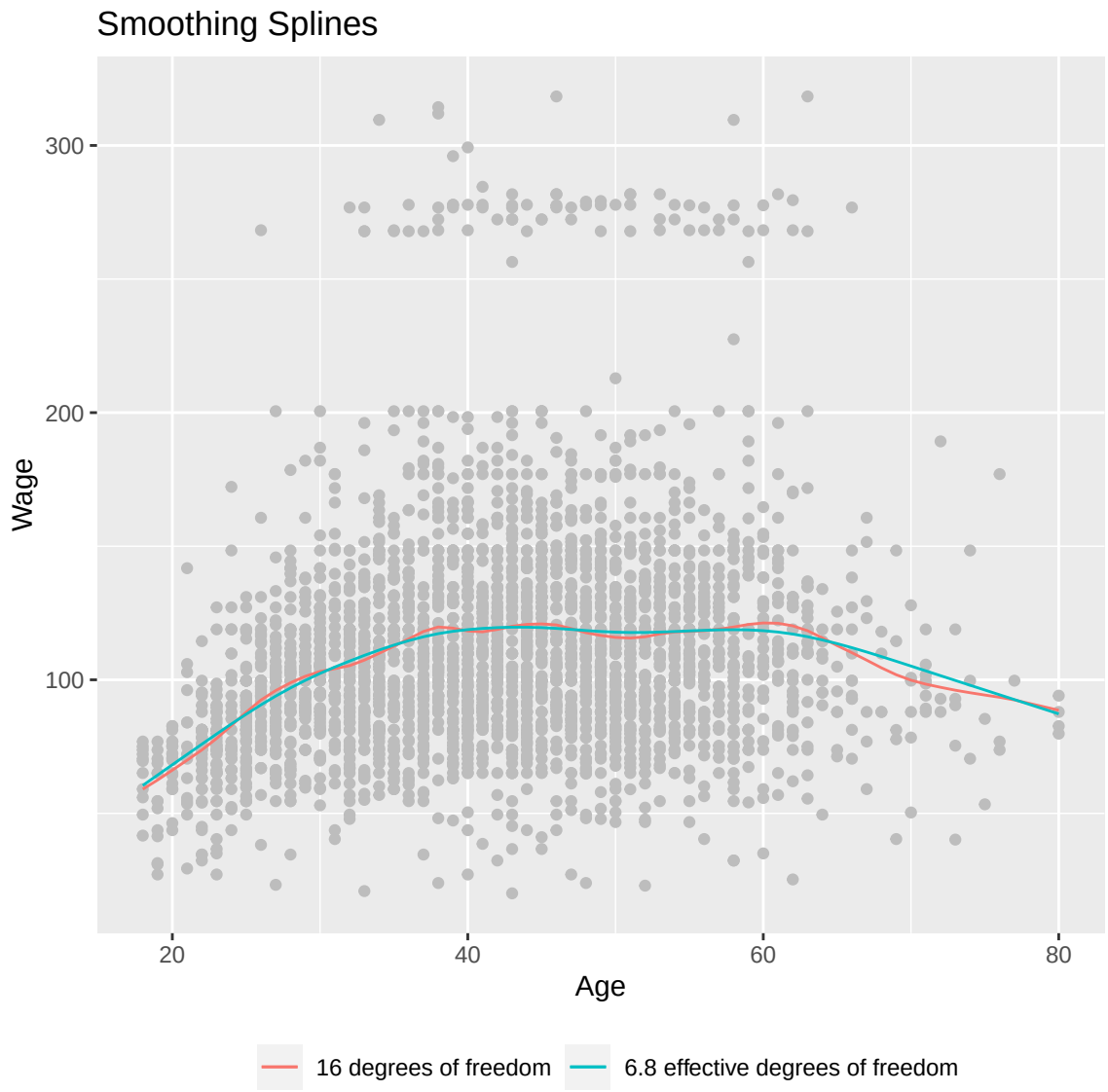


3.2 Smoothing Splines

```
# Fit 2 smoothing splines
fit_smooth = with(Wage, smooth.spline(age, wage, df = 16))
fit_smooth_cv = with(Wage, smooth.spline(age, wage, cv = TRUE))


# Plot the smoothing splines
ggplot() +
  ggtitle("Smoothing Splines") +
  xlab("Age")+
  ylab("Wage") +
  geom_point(data = Wage, aes(x = age, y = wage),colour="gray74") +
  geom_line(aes(x = fit_smooth$x, y = fit_smooth$y,
                color = "16 degrees of freedom"))  +
```

```
  geom_line(aes(x = fit_smooth_cv$x, y = fit_smooth_cv$y,
                color = "6.8 effective degrees of freedom")) +
  theme(legend.position = 'bottom')+
  labs(title = "Smoothing Splines", colour="")
```

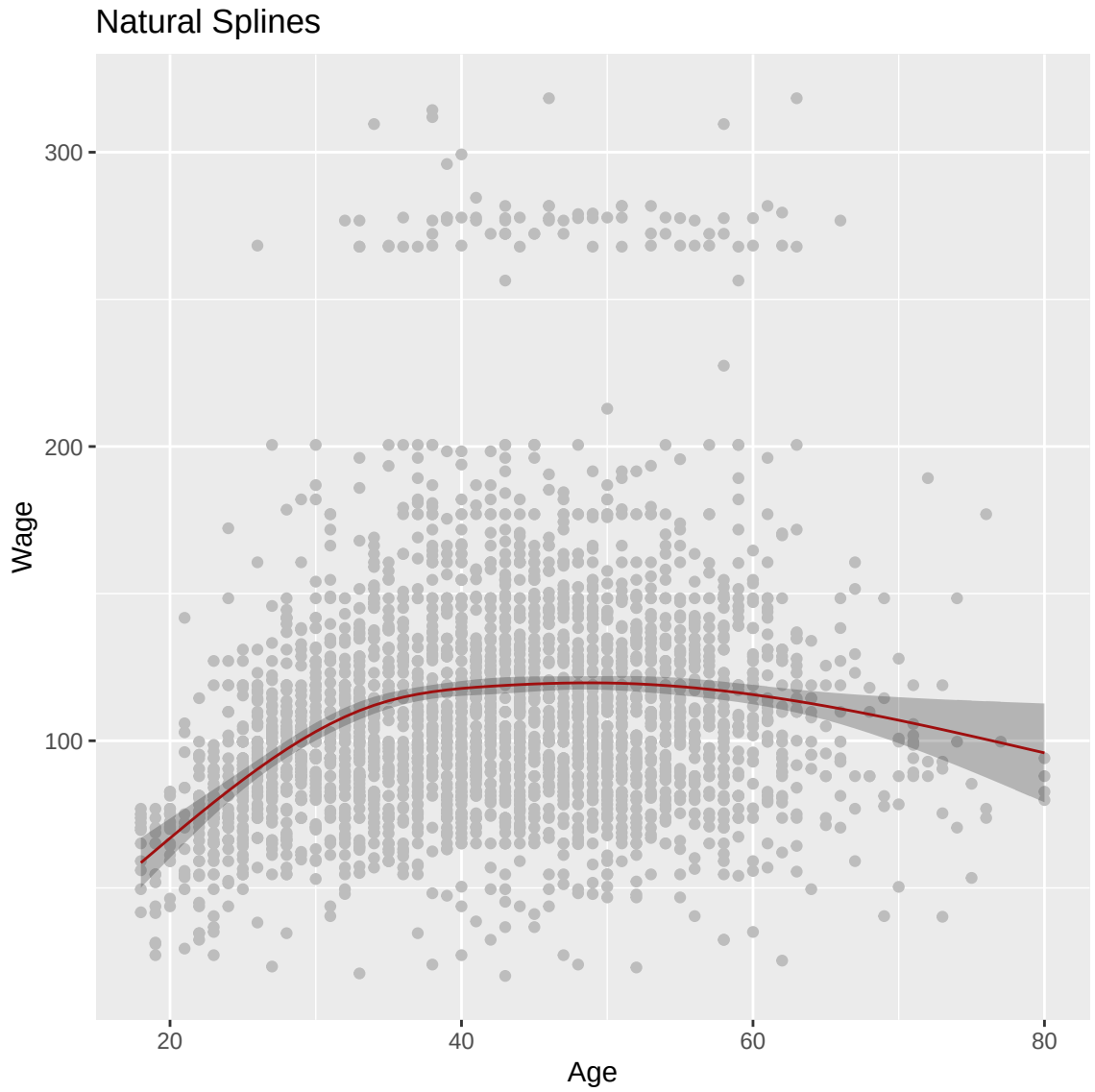


3.3 Natural Splines

```
natural.model = lm(wage~ns(age,df=4),data = Wage)
ns_predict = predict(natural.model, newdata= list(age = age_grid), se=TRUE)


# Compute error bands (2*SE)
se_bands = with(ns_predict, cbind("upper" = fit+2*se.fit,
                            "lower" = fit-2*se.fit))
```

```
# Plot the spline and error bands
ggplot() +
  ggtitle("Natural Splines") +
  xlab("Age")+
  ylab("Wage") +
  geom_point(data = Wage, aes(x = age, y = wage),colour="gray74") +
  geom_line(aes(x = age_grid, y = ns_predict$fit), color = "red3") +
  geom_ribbon(aes(x = age_grid,
                  ymin = se_bands[,"lower"],
                  ymax = se_bands[,"upper"]),
              alpha = 0.3) +
  xlim(agelims)
```

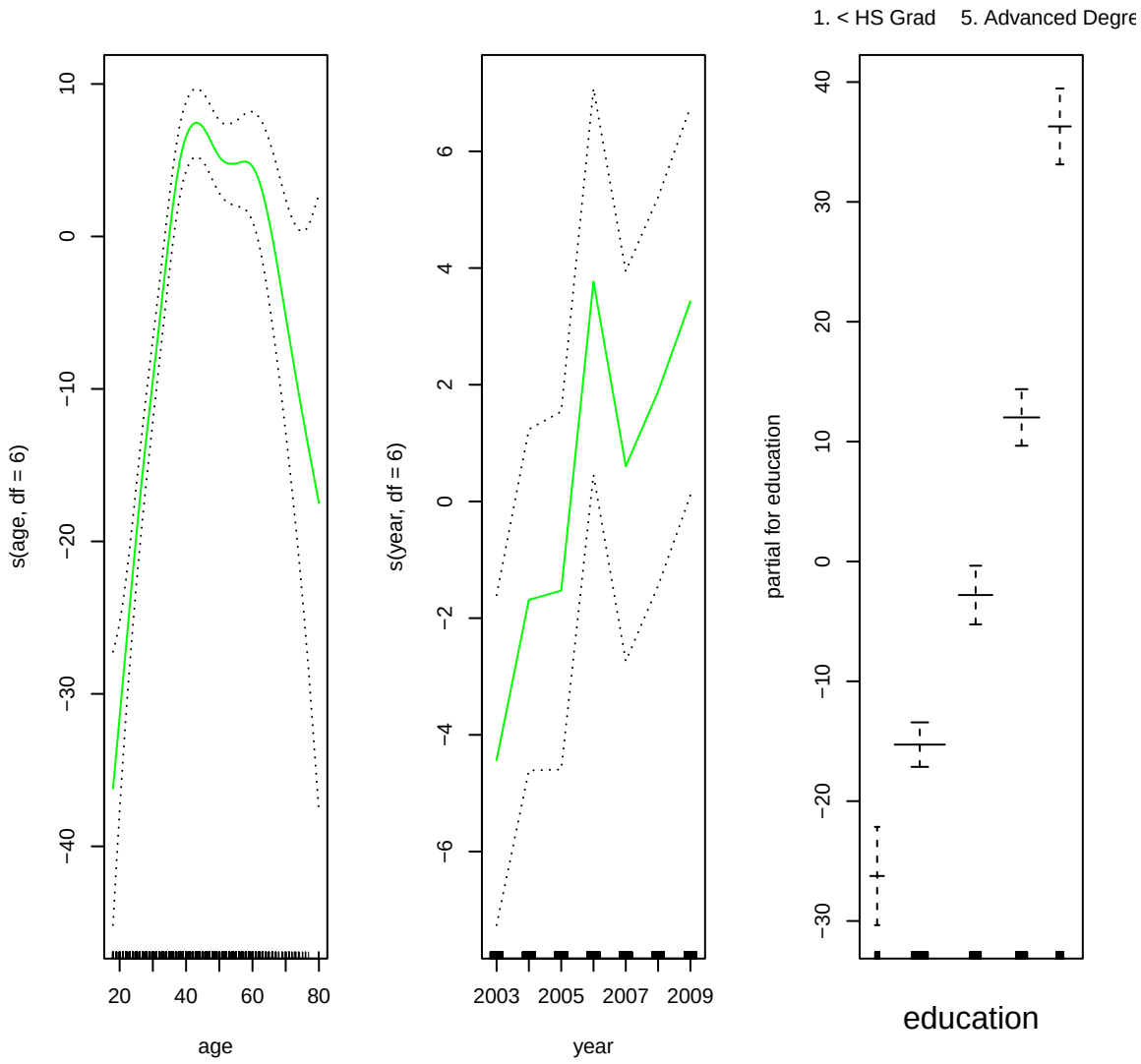


3.4 Using GAM Function

```
gam1<-gam(wage~s(age,df=6)+s(year,df=6)+education ,data = Wage)
par(mfrow=c(1,3)) #to partition the Plotting Window
summary(gam1)
```

```
##
## Call: gam(formula = wage ~ s(age, df = 6) + s(year, df = 6) + education,
##     data = Wage)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -119.89  -19.73   -3.28   14.27  214.45
##
## (Dispersion Parameter for gaussian family taken to be 1235.516)
##
##     Null Deviance: 5222086 on 2999 degrees of freedom
## Residual Deviance: 3685543 on 2983 degrees of freedom
## AIC: 29890.31
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## s(age, df = 6)     1  200717  200717 162.456 < 2.2e-16 ***
## s(year, df = 6)    1   22090   22090  17.879 2.425e-05 ***
## education          4 1069323  267331 216.372 < 2.2e-16 ***
## Residuals       2983 3685543    1236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                 Npar Df  Npar F  Pr(F)
## (Intercept)
## s(age, df = 6)        5 26.2089 <2e-16 ***
## s(year, df = 6)       5  1.0144 0.4074
## education
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Plotting the Model
plot(gam1,se = TRUE,col="green")
```



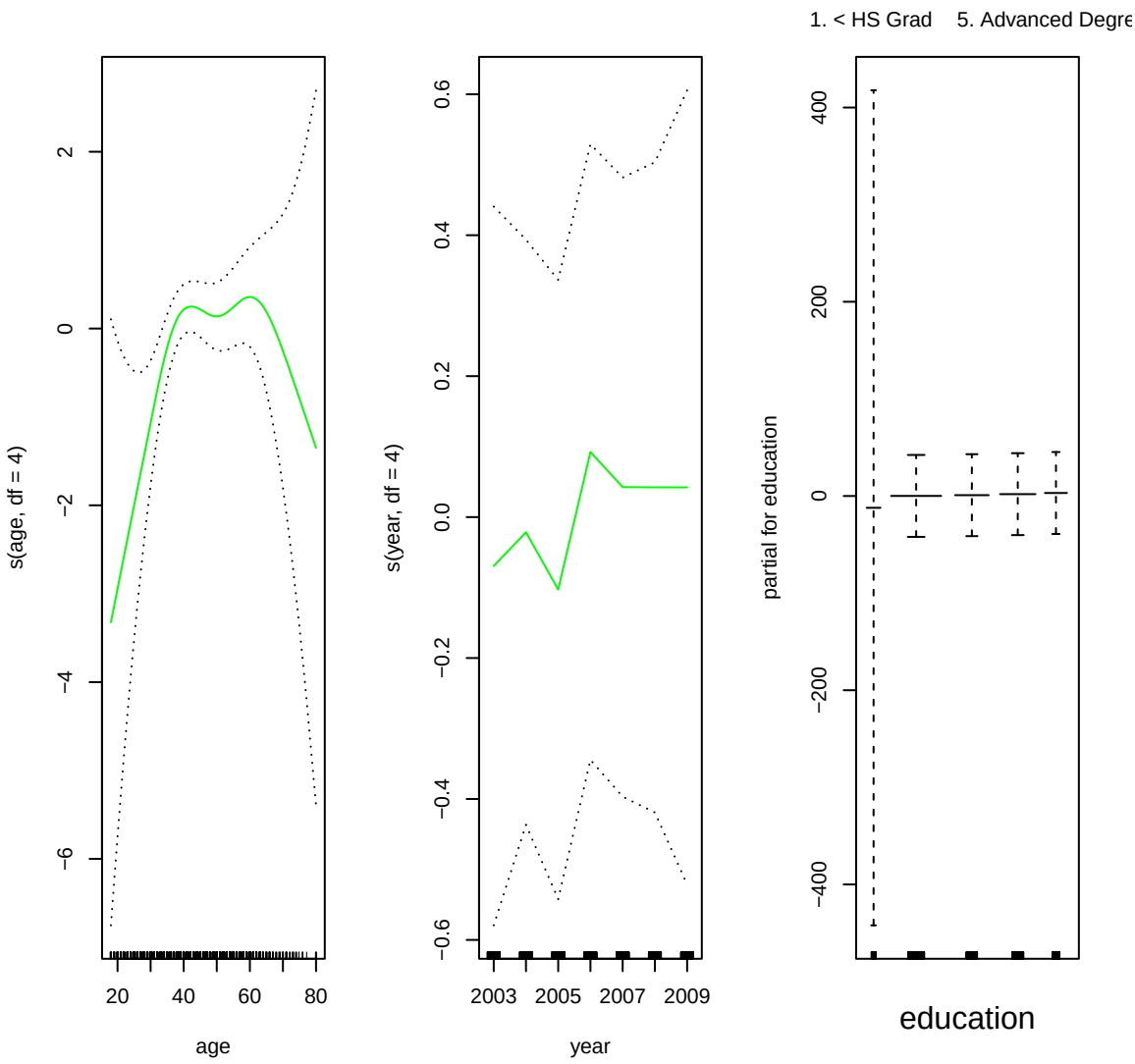

```
#se stands for standard error Bands
```

From the above plots:
1. wages increases with age then decreases at around 60.
2. There is decrease on salary at around year 2007 or 2008.
3. Wages increases with the level of education

3.5 Logistic Regression Using GAM

We can have the logistic model to predict a person can earn more than or less than 250 based on the age,year and education.

```
logitgam1<-gam(I(wage > 250) ~ s(age,df=4) + s(year,df=4) + education ,data=Wage,family=binomial)
par(mfrow=c(1,3))
plot(logitgam1,se=T,col="green")
```


s(age, df = 4)
2
0
−2
−4
−6
20
40
60
80
age
s(year, df = 4)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
2003
2005
2007
2009
year
1. < HS Grad
5. Advanced Degre
partial for education
400
200
0
−200
−400
education