Analysis of Wage Infomration Using GAMs

Gap Kim

INTRODUCTION

In this report, generalized additive models (GAMs) have been used to analyze ‘Wage’ dataset from ISLR package. The dataset contains information from 3000 male individuals from the mid-Atlantic region. Two GAMs have been built to predict quantitative response ‘logwage’ and to predict whether an individual earns more than $250,000 or not. GAMs were built with smoothing splines for the quantitative predictors, ‘age’ and ‘year’, and the categorical predictors were considered for the model. A generalized cross validation was performed to obtain the optimal degrees of freedom for the smoothing spline. Based on hypothesis testing with ANOVA, a best GAM has been suggested each for predicting quantitative response ‘logwage’ and predicting whether an individual earns more than $250,000 or not. The selected predictors were different depending on the type of prediction selected. Based on the developed best models, traits of individuals who have higher wages have been identified.

EXPLORATORY DATA ANALYSIS

The dataset contains 3000 individuals with information on the following 11 variables.

- year: Year that wage information was recorded
- age: Age of worker
- maritl: A factor with levels 1. Never Married 2. Married 3. Widowed 4. Divorced and 5. Separated indicating marital status
- race: A factor with levels 1. White 2. Black 3. Asian and 4. Other indicating race
- education: A factor with levels 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree indicating education level
- region: Region of the country (mid-atlantic only)
- jobclass: A factor with levels 1. Industrial and 2. Information indicating type of job
- health: A factor with levels 1. <=Good and 2. >=Very Good indicating health level of worker
- health_ins: A factor with levels 1. Yes and 2. No indicating whether worker has health insurance
- logwage: Log of workers wage
- wage: Workers raw wage

The ‘region’ variable only had a single factor, mid-Atlantic, and therefore, was eliminated from the analysis. The boxplots of ‘logwage’ has been plotted for all 6 categorical predictors in Figure 1. From Figure 1, a linear pattern is observed between ‘logwage’ and increasing ‘education’. It also seems having ‘health_ins’ is associate with higher ‘logwage’.

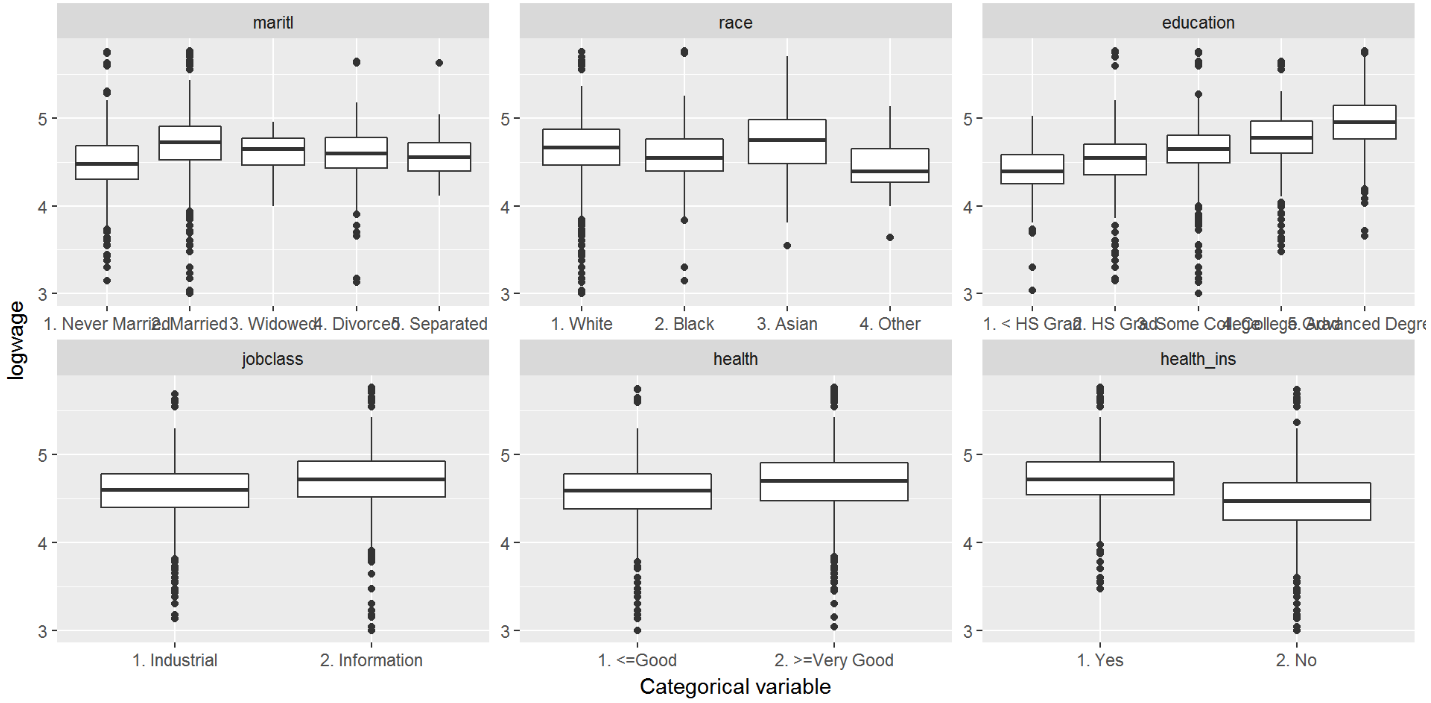

Figure 1: Boxplots of ‘logwage’ for categorical variables

Expanding upon the observations made in Figure 1, scatterplots of ‘logwage’ and ‘age’ have been plotted in Figure 2 with information of ‘health_ins’ and ‘eduction’ level. On the top figure, individuals with ‘health_ins’ tend to have higher wage than those who do not. On the bottom figure, individuals with higher education level earned more.

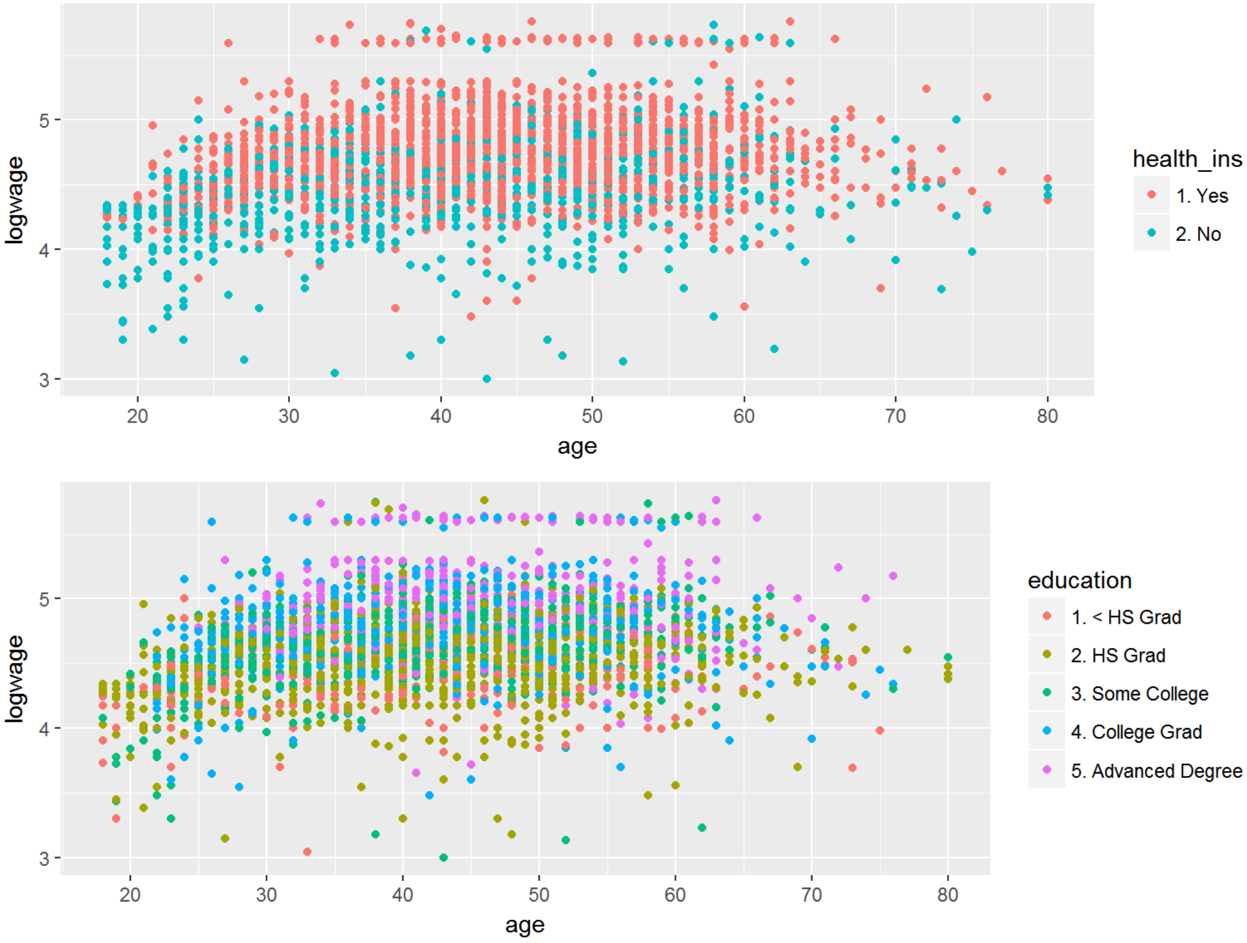

Figure 2: Scatterplot of ‘wage’ vs. ‘age’ by ‘health_ins’ and ‘education’

ANALYSIS USING NONLINEAR METHODS

Prediction of ‘logwage’ response

GAM with smoothing splines

To apply non-linear function to the predictors, generalized additive modeling (GAM) approach has been used to predict the response ‘logwage’. Smoothing splines have been applied to the two quantitative predictors, ‘age’ and ‘year’. To find optimal degrees of freedom (DF) for the smoothing splines, a generalized cross validation scheme has been performed with following results:

```
Optimal DF of the smoothing spline for 'age':  7.567
```

```
Optimal DF of the smoothing spline for 'year':  2.804
```

Using the acquired optimal DFs for ‘age’ and ‘year’, hypothesis testing using ANOVA has been performed to identify significant predictors. Eight separate models were tested where each model included additonal variables in the following order (from left to right):

$$logwage = s(age, df = 7.567) + s(year, df = 2.804) + education + health\_ins + maritl + health + jobclass + race$$

```
Analysis of Deviance Table

Model 1: logwage ~ s(age, df = 7.567)
Model 2: logwage ~ s(age, df = 7.567) + s(year, df = 2.804)
Model 3: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education
Model 4: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education +
    health_ins
Model 5: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education +
    health_ins + maritl
Model 6: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education +
    health_ins + maritl + health
Model 7: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education +
    health_ins + maritl + health + jobclass
Model 8: logwage ~ s(age, df = 7.567) + s(year, df = 2.804) + education +
    health_ins + maritl + health + jobclass + race
  Resid. Df Resid. Dev  Df Deviance      F  Pr(>F)    
1      2991        327                                
2      2989        324 2.8      2.5  11.99 2.0e-07 ***
3      2985        255 4.0     68.7 226.97 < 2e-16 ***
4      2984        236 1.0     19.9 262.53 < 2e-16 ***
5      2980        228 4.0      7.8  25.70 < 2e-16 ***
6      2979        226 1.0      2.0  25.94 3.8e-07 ***
7      2978        226 1.0      0.3   3.93   0.048 *  
8      2975        225 3.0      0.5   2.25   0.080 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA ouptut is shown above for the 8 GAMs. The p-value for each model indicates that adding all subsequent predictors except the ‘race’ (the last predictor added) are statistically significant in predicting ‘logwage’. Hence the model with following 7 predictors has been selected as the best GAM in predicting ‘logwage’:

$$logwage = s(age, df = 7.567) + s(year, df = 2.804) + education + health\_ins + maritl + health + jobclass$$

The partial regression plots for each predictor of the best model are shown in Figre. 3.

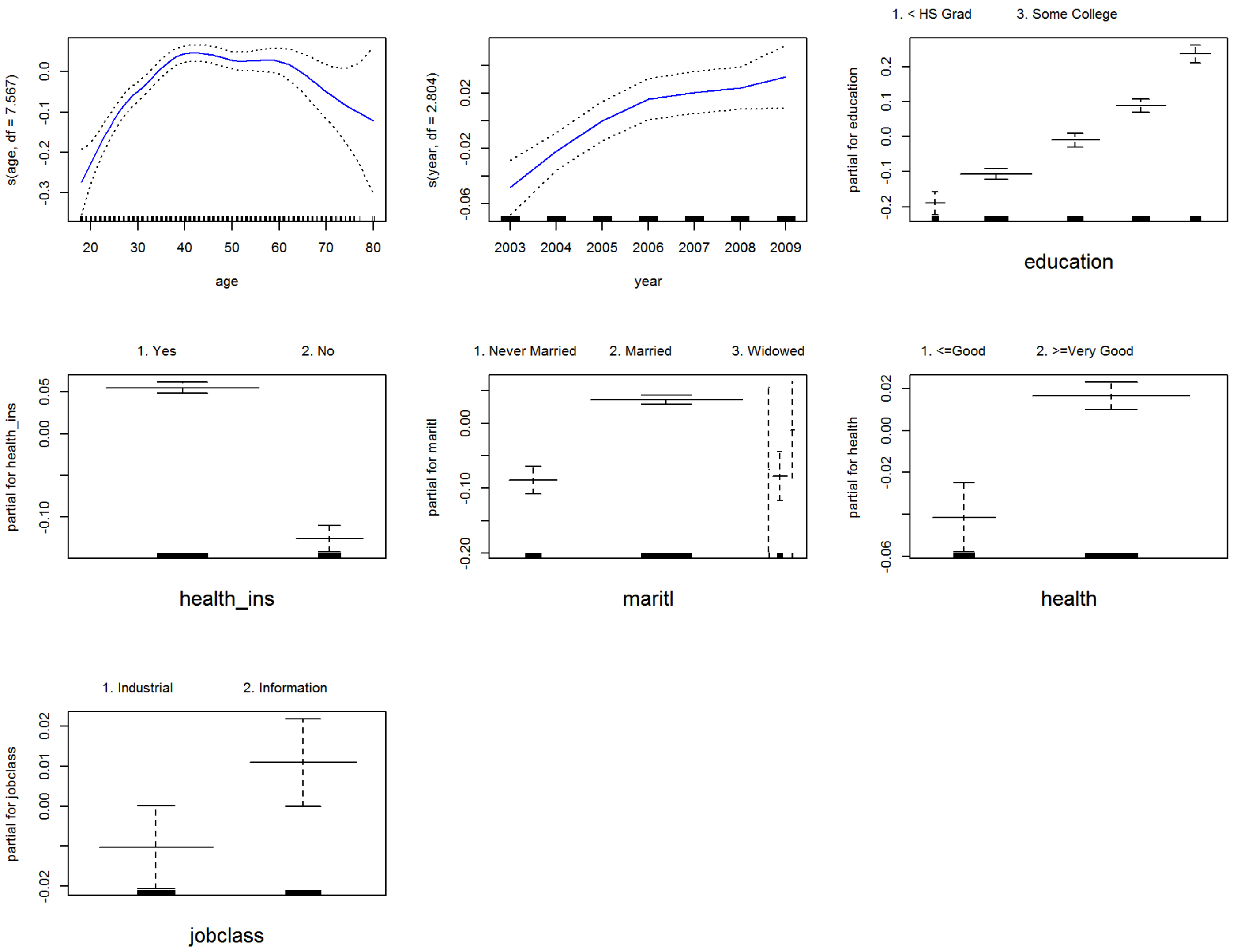

Figure 3: Partial regression plots from best GAM

The partial regression plot of ‘age’ using smoothing spline captures the shape of high wage individuals for ages between 40 and 60; however, the standard error becomes rather high for low and high ages. The partial regression plot of ‘year’ shows almost a linear increase in ‘logwage’ with increasing ‘year’. The partial regression plot for ‘education’ tells that the wage is higher for individuals with higher degree. It is also apparent that individuals with health insurance and with very good health condition earn higher wages that those who do not. Also, married individuals earn higher wages than those who are never married.

Diagonostic plots

Diagnostic plots have been examined if there are issues with the model. First, residual versus fitted values have been plotted in Figure 4. Overall, the distribution of residuals look randomly distributed below and above 0 supporting linearity of the terms. Also, residuals form a band supporting constant variance of the error terms. It is noted that a band of residuals with green color is observed for those individuals who earn above $250,000.

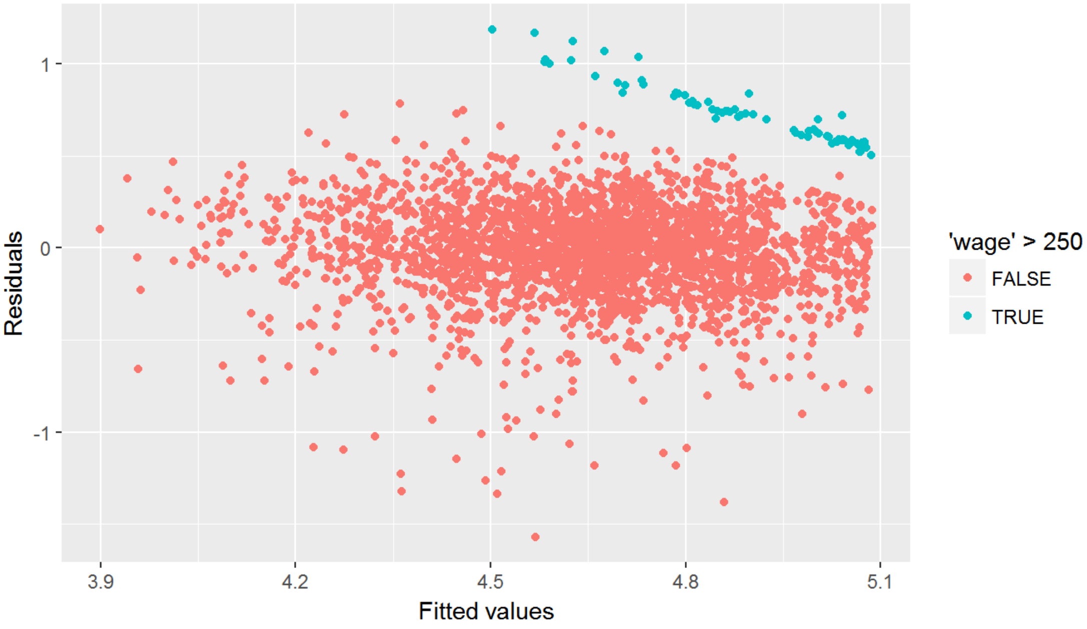

Figure 4: Residual versus fits plot

Residual versus predictor plots have been provided for ‘age’ and ‘year’ in Figure 5. The plots do not show much concern for lack of fit or nonconstant error variance for the predictors.

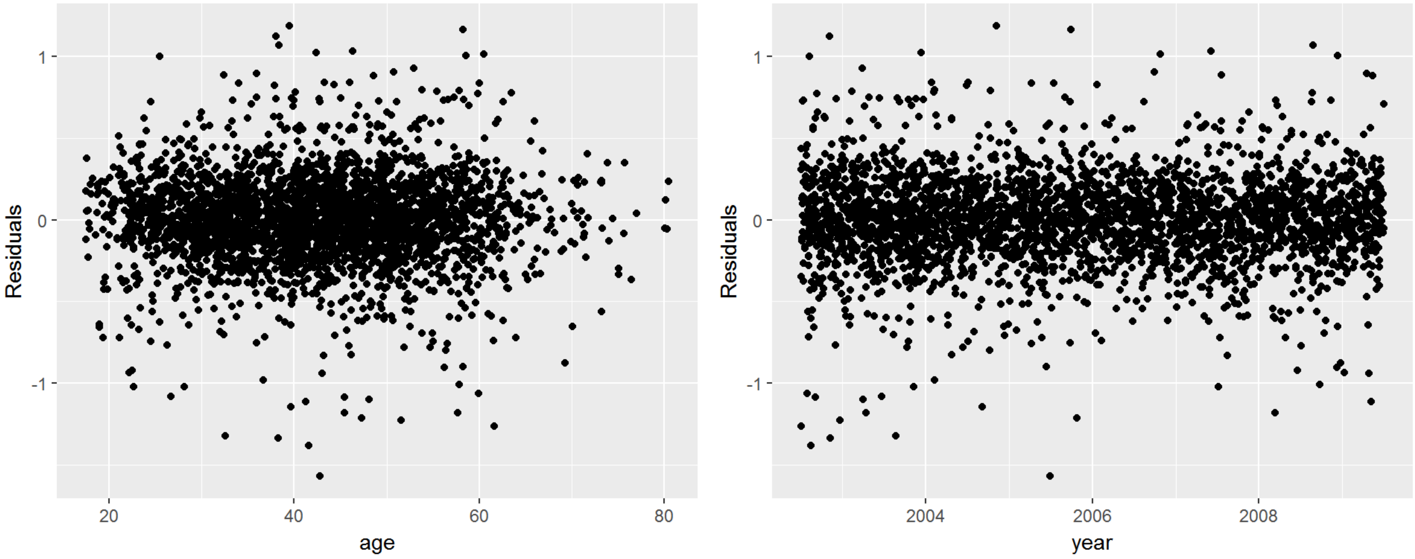

Figure 5: Residual versus predictors

The Q-Q plot and histogram of residuals are provided in Figure 6. It seems the distribution of residuals are heavy-tailed. What this means is that there are more frequent amount of large residuals compared with normal distribution.

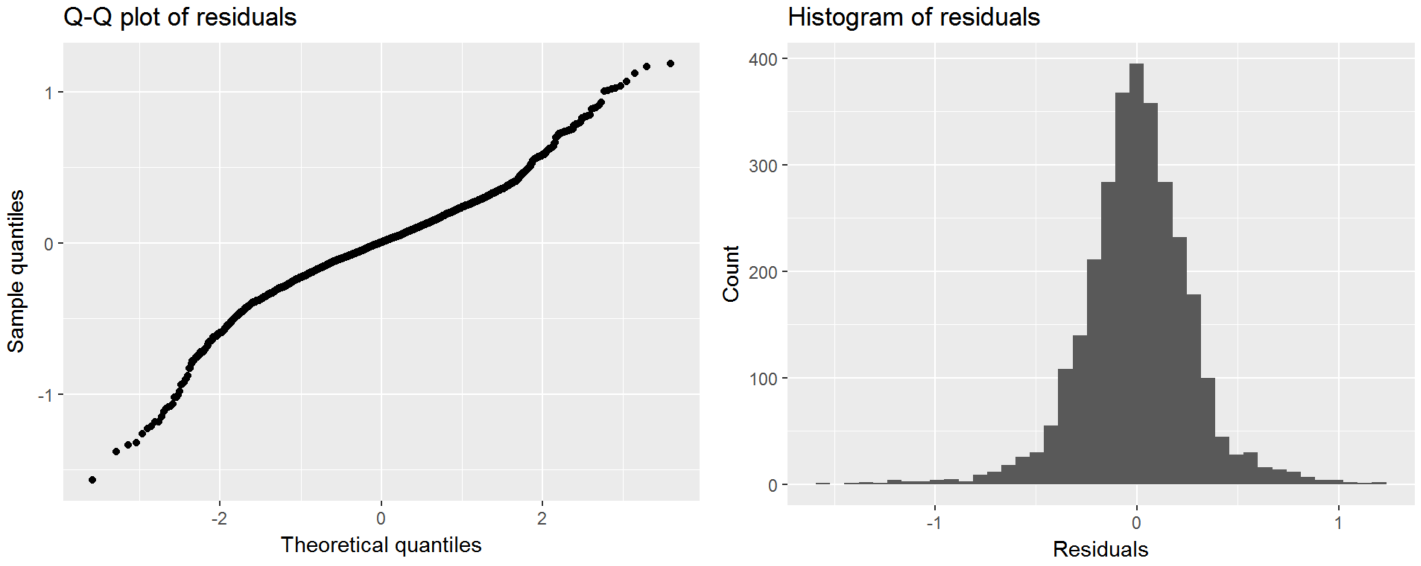

Figure 6: Q-Q plot and histogram of residuals

The heavy-tailed distibution of residuals can also be undertood from the fitted versus response plot in Figure. 7. The linear relationship between fitted values and response is apparent when logwage values are beween 4 and 5. For high and low values of ‘logwage’, however, the ‘logwage’ values deviate from the linear trend and the residuals become large.

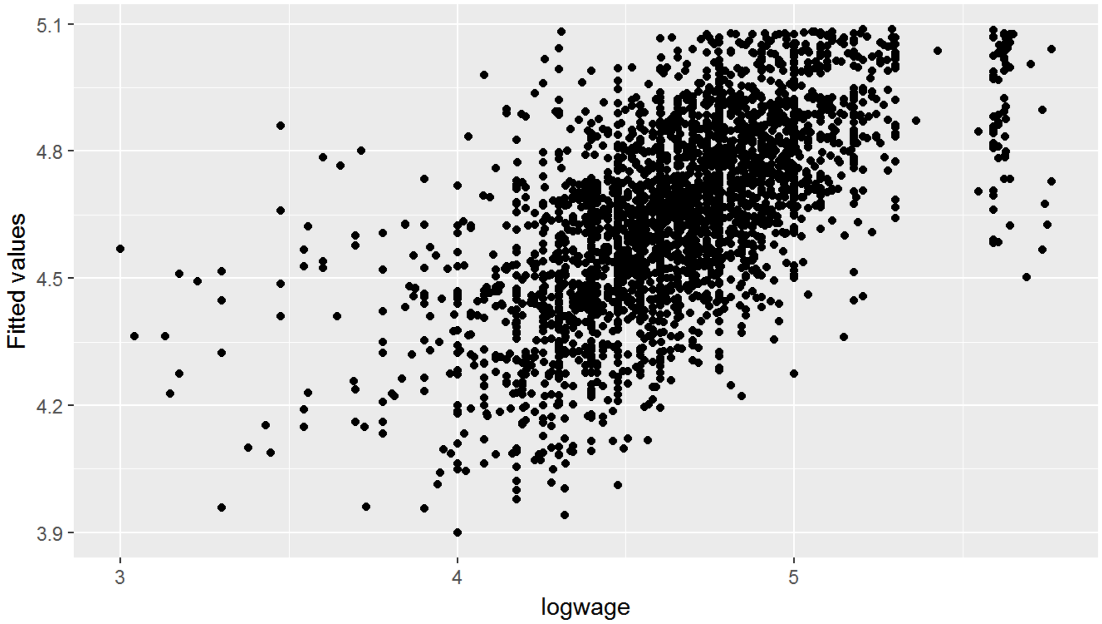

Figure 7: Fitted versus Response (logwage)

The performance of GAMs used in predicting ‘logwage’ is summarized in Table 1. The residual deviance of each model decreased as additional predictors are included in the model although very little difference is found on the last three models. The Akaike information criterion (AIC) also shows a similar trend where very little change is found for the last two models (Gam7 and Gam8). The results agree with best model selected by hypothesis testing using ANOVA.

Table 1: The performance of GAMs used in predicting ‘logwage’

	Gam1	Gam2	Gam3	Gam4	Gam5	Gam6	Gam7**	Gam8
Predictor								
year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
age	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
education	No	No	Yes	Yes	Yes	Yes	Yes	Yes
health_ins	No	No	No	Yes	Yes	Yes	Yes	Yes
maritl	No	No	No	No	Yes	Yes	Yes	Yes
health	No	No	No	No	No	Yes	Yes	Yes
jobclass	No	No	No	No	No	No	Yes	Yes

** Selected best model

Table 1: The performance of GAMs used in predicting ‘logwage’

	Gam1	Gam2	Gam3	Gam4	Gam5	Gam6	Gam7**	Gam8
race	No	No	No	No	No	No	No	Yes
Metric								
AIC	1880.6	1862.8	1156.4	915.5	822.8	798.9	796.9	796.1
Res_deviance	326.7	324.1	255.5	235.6	227.8	225.8	225.6	225

** Selected best model

Prediction of binary ‘wage’ data

From Figure 2, it is apparent that the wage data can be treated as a binary categorical variable depending on whether an individual earns above or under $250,000. The generlized additive model framework has been used to predict the binary ‘wage’ response.

GAM with smoothing splines

Using the optimal DFs for ‘age’ and ‘year’ found in the previous section, hypothesis testing with ANOVA has been performed to identify the significant predictors.

Eight separate models have been tested where each model includes an additonal variable in the following order:

$$wage(>250) = education + maritl + jobclass + s(age, df = 7.567) + health + health\_ins + race + s(year, df = 2.804)$$

Note the difference in the order from the quantitative response prediction. A preanalysis had been performed to determine the level of significance of each predictor on the prediction of the binary ’wage’response. The results from ANOVA are shown below:

```
Analysis of Deviance Table

Model 1: I(wage > 250) ~ education
Model 2: I(wage > 250) ~ education + maritl
Model 3: I(wage > 250) ~ education + maritl + jobclass
Model 4: I(wage > 250) ~ education + maritl + jobclass + s(age, df = 7.567)
Model 5: I(wage > 250) ~ education + maritl + jobclass + s(age, df = 7.567) +
    health
Model 6: I(wage > 250) ~ education + maritl + jobclass + s(age, df = 7.567) +
    health + health_ins
Model 7: I(wage > 250) ~ education + maritl + jobclass + s(age, df = 7.567) +
    health + health_ins + race
Model 8: I(wage > 250) ~ education + maritl + jobclass + s(age, df = 7.567) +
    health + health_ins + race + s(year, df = 2.804)
  Resid. Df Resid. Dev   Df Deviance    F Pr(>F)  
1      2995        622                            
2      2991        609 4.00    13.24 3.31  0.010 *
3      2990        604 1.00     5.07 5.07  0.024 *
4      2982        589 7.57    14.54 1.92  0.056 .
5      2981        585 1.00     3.63 3.62  0.057 .
6      2980        584 1.00     1.83 1.83  0.176  
7      2977        582 3.00     1.60 0.53  0.660  
8      2975        582 2.80     0.38 0.13  0.930  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Models 1, 2 and 3 are significant, and the next two models, 4 and 5, are marginally significant. Predictors, ‘health_ins’, ‘race’ and ‘year’ are insignficant when a model already contains ‘education’, ‘maritl’, ‘jobclass’, ‘age’ and ‘health’. Thus, Model 5 has been selected as the best model for predicting the binary ‘wage’ response. A summary of models used in predicting the binary ‘wage’ response and their performance is provided in Table 2. The AIC of the GamB5 (Model 5) is the lowest, and thus, agrees with best model selected from hypothesis testing.

Table 2: The performance of GAMs used in predicting binary ‘wage’ response

	GamB1	GamB2	GamB3	GamB4	GamB5**	GamB6	GamB7	GamB8
Predictor								
education	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

** Selected best model

Table 2: The performance of GAMs used in predicting binary ‘wage’ response

	GamB1	GamB2	GamB3	GamB4	GamB5**	GamB6	GamB7	GamB8
maritl	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
jobclass	No	No	Yes	Yes	Yes	Yes	Yes	Yes
age	No	No	No	Yes	Yes	Yes	Yes	Yes
health	No	No	No	No	Yes	Yes	Yes	Yes
health_ins	No	No	No	No	No	Yes	Yes	Yes
race	No	No	No	No	No	No	Yes	Yes
year	No	No	No	No	No	No	No	Yes
Metric								
AIC	632	626.7	623.7	624.2	622.6	622.8	627.2	632.4
Res_deviance	622	608.7	603.7	589.1	585.5	583.7	582.1	581.7

** Selected best model

Upon plotting the partial classification for each of the predictor as shown in Figure 8, ‘education’ and ‘maritl’ plots seem suspicious with very wide standard errors. Due to the large standard error bars, the influences of factor levels in ‘education’ and ‘maritl’ are not noticeable.

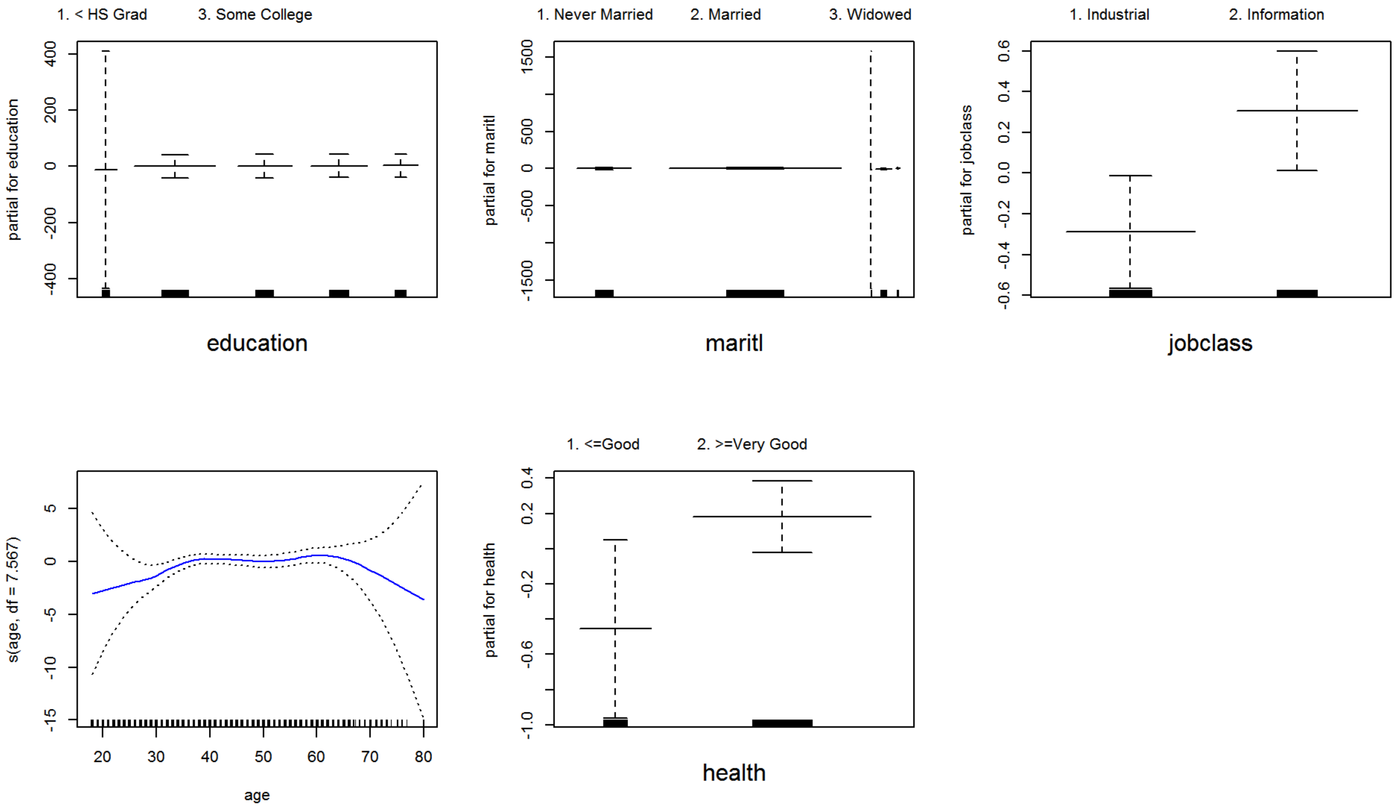

Figure 8: Partial classification plots of each predictor resulting from GamB5(Model 5)

It turns out that there is no one who earns higher than $250,000 with a degree less than HS Grad or widowed. There are only two individuals who earn higher than $250,000 who are divorced; and only a single person who earns higher than $250,000 and is separated.

After eliminating datapoints for these empty factors (or has only 1 or 2 individuals), the model has been refit with 2486 datapoints. The results are shown in Figure 9. The partial classification plot of ‘education’ shows that a person is more likely to earn more the $250,000 with higher education level. It is more likely for individuals who are married, work in information related job, and have a very good or higher health condition earn more than $250,000.

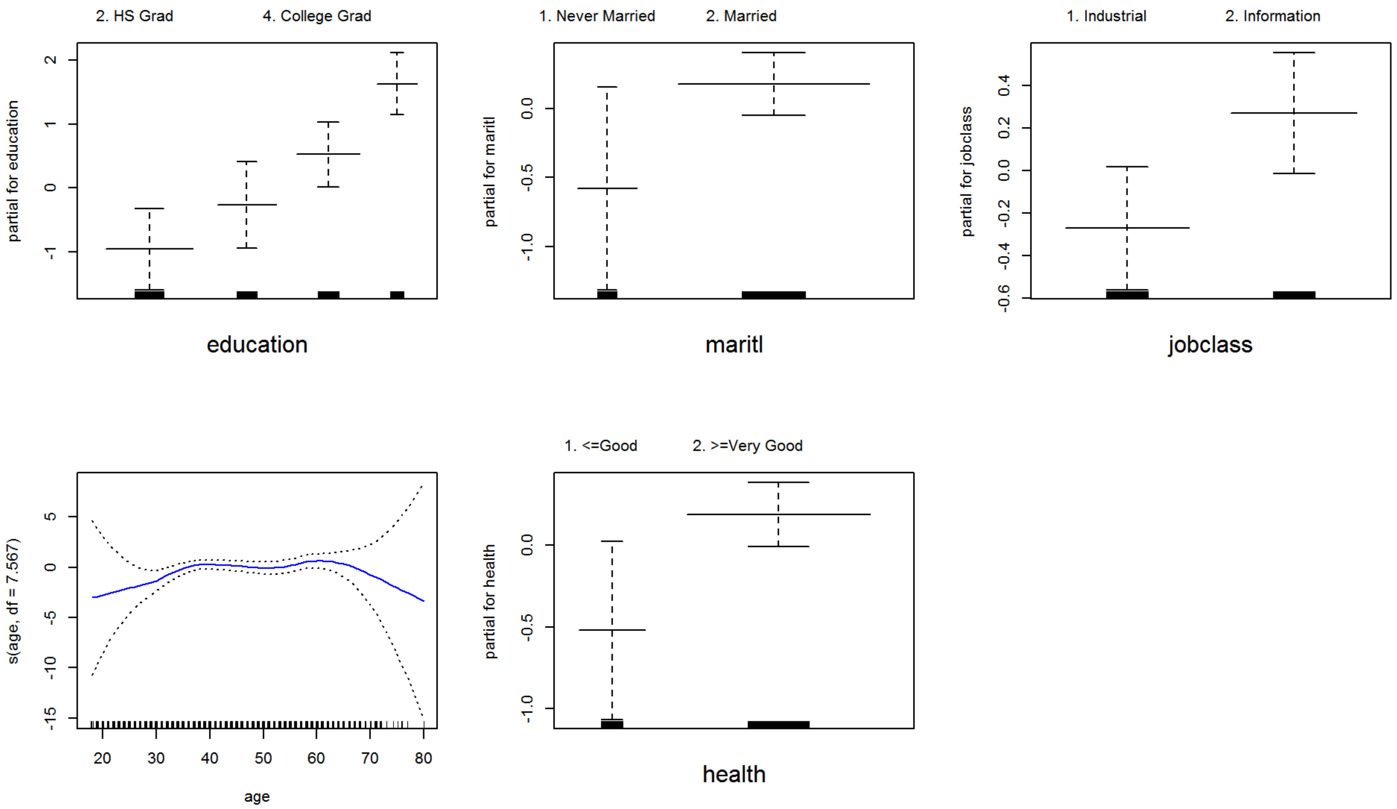

Figure 9: Partial classification plots of each predictor resulting from Model 5 after dropping null factors

Using the developed model, the posterior probability of an individual making more than $250,000 is calculated to be 0.1255 given that an individual is 50 years old and married with excellent health condition, and has an advanced degree working in information job.

CONCLUSIONS

Generalized additive models (GAMs) have been used to analyze ‘Wage’ dataset from ISLR package. The dataset contains information from 3000 male individuals from the mid-Atlantic region. Two GAMs have been built to predict quantitative response ‘logwage’ and to predict whether an individual earns more than $250,000 or not.

GAMs were built with smoothing splines for the quantitative predictors, ‘age’ and ‘year’, where optimal degrees of freedoms were determined as 7.567 and 2.804 for ‘age’ and ‘year’, respectively. A hypothesis testing with ANOVA was used to identify signficant predictors. The best model in predicting the ‘logwage’ included predictors, ‘age’, ‘year’, ‘education’, ‘health_ins’, ‘maritl’, ‘health’, and ‘jobclass’ with residual deviance of 225.6. The best model in predicting the binary ‘wage > $250,000’ included predictors ‘education’, ‘maritl’, ‘jobclass’, ‘age’, and ‘health’ with residual deviance of 585.5.

Based on the partial regression and classification plots from the best models, we can conclude that individuals are more likely to have higher wages who have following traits:

- between ages 40 and 60 years old
- higher the education level
- married
- have very good or excellent health condition
- work in information related jobs than those in industrial
- have health insurance

APPENDIX

All R codes used in producing the results are included below:

```
##############################
### Initial Setup

knitr::opts_chunk$set(comment=NA, echo=FALSE, warning=FALSE, message=FALSE,
                      fig.align="center")
options(digits=4)
rm(list=ls())

library(ISLR)
data(Wage)

#table(Wage$region)

Wage = Wage[, -6]

#sum(is.na(Wage))
#str(Wage)

library(ggplot2)
library(reshape2)

Wage.cat = Wage[,c(3:9,10)]
Wage.cat.melt = melt(Wage.cat[,-8], id.vars = "logwage")

ggplot(data=Wage.cat.melt) +
   geom_boxplot(aes(x=value, y=logwage)) +
   facet_wrap(~variable, scale="free") +
   labs(x="Categorical variable")

library(gridExtra)

p1 = ggplot(data=Wage) +
   geom_point(aes(x=age, y=logwage, color=health_ins)) +
   scale_x_continuous(breaks=seq(20,80, by=10))

p2 = ggplot(data=Wage) +
   geom_point(aes(x=age, y=logwage, color=education)) +
   scale_x_continuous(breaks=seq(20,80, by=10))

grid.arrange(p1, p2, ncol=1)

attach(Wage)
agelims = range(age)
age.grid = seq(from=agelims[1], to=agelims[2])

smooth.age = smooth.spline(x=age, y=logwage, cv=FALSE)
cat("Optimal DF of the smoothing spline for 'age': ", smooth.age$df)

smooth.year = smooth.spline(x=year, y=logwage, cv=FALSE)
cat("Optimal DF of the smoothing spline for 'year': ", smooth.year$df)

library(gam)

gam.m1 = gam(logwage ~ s(age, df=7.567), data=Wage)
gam.m2 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804), data=Wage)
gam.m3 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education, data=Wage)
gam.m4 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education + health_ins, data=Wage)
gam.m5 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education + health_ins +
                maritl, data=Wage)
gam.m6 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education + health_ins +
                maritl + health, data=Wage)
gam.m7 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education + health_ins +
                maritl + health + jobclass, data=Wage)
gam.m8 = gam(logwage ~ s(age, df=7.567) + s(year, df=2.804) + education + health_ins +
                maritl + health + jobclass + race , data=Wage)

anova(gam.m1, gam.m2, gam.m3, gam.m4, gam.m5, gam.m6, gam.m7, gam.m8, test="F")

par(mfrow = c(3,3))
plot(gam.m7, se=T, col="blue")

ggplot() +
   geom_point(aes(x=gam.m7$fitted.values, y=gam.m7$residuals, color=as.factor(I(wage>250)))) +
   labs(x="Fitted values", y="Residuals", color="'wage' > 250")

library(gridExtra)

dd1 = ggplot() +
   geom_jitter(aes(x=Wage$age , y=gam.m7$residuals), width=0.5) +
   labs(x="age", y="Residuals")

dd2 = ggplot() +
   geom_jitter(aes(x=Wage$year , y=gam.m7$residuals), width=0.5) +
   labs(x="year", y="Residuals")

grid.arrange(dd1, dd2, ncol=2)
dd3 = ggplot() +
   geom_qq(aes(sample=gam.m7$residuals)) +
   labs(x="Theoretical quantiles", y="Sample quantiles") +
   ggtitle("Q-Q plot of residuals")

dd4 = ggplot() +
   geom_histogram(aes(x=gam.m7$residuals), bins=40) +
   labs(x="Residuals", y="Count") +
   ggtitle("Histogram of residuals")

grid.arrange(dd3, dd4, ncol=2)


ggplot() +
   geom_point(aes(x=Wage$logwage, y=gam.m7$fitted.values)) +
   labs(x="logwage", y="Fitted values")

library(htmlTable)

rname = c("year","age","education","health_ins","maritl","health","jobclass","race")
AIC = c(gam.m1$aic, gam.m2$aic, gam.m3$aic, gam.m4$aic, gam.m5$aic, gam.m6$aic,
        gam.m7$aic, gam.m8$aic)
res.deviance = c(gam.m1$deviance, gam.m2$deviance, gam.m3$deviance, gam.m4$deviance,
                 gam.m5$deviance, gam.m6$deviance, gam.m7$deviance, gam.m8$deviance)

feature.m = matrix(rep("No", 8*8), nrow=8)
for(i in c(1:8)) {
   feature.m[i,i:8] = "Yes"
}

rownames(feature.m) = rname
colnames(feature.m) = c("Gam1","Gam2","Gam3","Gam4","Gam5","Gam6","Gam7**","Gam8")
feature.m = rbind(feature.m, round(AIC,1), round(res.deviance,1))

rownames(feature.m)[9:10] = c("AIC", "Res_deviance")

htmlTable(feature.m,
          caption="Table 1: The performance of GAMs used in predicting 'logwage'",
          rgroup = c("Predictor", "Metric"),
          n.rgroup = c(8,2),
          tfoot = paste("** Selected best model"),
          css.cell = "width:80px;")
```

```
gam.b1 = gam(I(wage>250) ~ education, family=binomial)
gam.b2 = gam(I(wage>250) ~ education + maritl, data=Wage, family=binomial)
gam.b3 = gam(I(wage>250) ~ education + maritl + jobclass, data=Wage, family=binomial)
gam.b4 = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567), data=Wage, family=binomial)
gam.b5 = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567) + health,
             data=Wage, family=binomial)
gam.b6 = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567) + health +
                health_ins, data=Wage, family=binomial)
gam.b7 = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567) + health +
                health_ins + race, data=Wage, family=binomial)
gam.b8 = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567) + health +
                health_ins + race + s(year, df=2.804), data=Wage, family=binomial)

anova(gam.b1, gam.b2, gam.b3, gam.b4, gam.b5, gam.b6, gam.b7, gam.b8, test="F")

rname = c("education","maritl","jobclass","age","health","health_ins","race","year")

AIC = c(gam.b1$aic, gam.b2$aic, gam.b3$aic, gam.b4$aic, gam.b5$aic, gam.b6$aic,
        gam.b7$aic, gam.b8$aic)
res.deviance = c(gam.b1$deviance, gam.b2$deviance, gam.b3$deviance, gam.b4$deviance,
                 gam.b5$deviance, gam.b6$deviance, gam.b7$deviance, gam.b8$deviance)

feature.m = matrix(rep("No", 8*8), nrow=8)
for(i in c(1:8)) {
   feature.m[i,i:8] = "Yes"
}

rownames(feature.m) = rname
colnames(feature.m) = c("GamB1","GamB2","GamB3","GamB4","GamB5**","GamB6","GamB7","GamB8")
feature.m = rbind(feature.m, round(AIC,1), round(res.deviance,1))

rownames(feature.m)[9:10] = c("AIC", "Res_deviance")

htmlTable(feature.m,
          caption = "Table 2: The performance of GAMs used in predicting binary 'wage' response",
          rgroup = c("Predictor", "Metric"),
          n.rgroup = c(8,2),
          tfoot = paste("** Selected best model"),
          css.cell = "width:80px;")

par(mfrow = c(2,3))
plot(gam.b5, se=T, col="blue")
check1 = sum((Wage$education == "1. < HS Grad") & (Wage$wage > 250))
check2 = sum((Wage$maritl == "3. Widowed") & (Wage$wage > 250))
check3 = sum((Wage$maritl == "4. Divorced") & (Wage$wage > 250))
check4 = sum((Wage$maritl == "5. Separated") & (Wage$wage > 250))

no.HS = which(Wage$education != "1. < HS Grad")
never.married = which(Wage$maritl == "1. Never Married")
married = which(Wage$maritl == "2. Married")
marital = union(never.married, married)

id = intersect(no.HS, marital)
numb.id = length(id)

Wage.rev = Wage[id,]

gam.b5.rev = gam(I(wage>250) ~ education + maritl + jobclass + s(age, df=7.567) + health,
             data=Wage.rev, family=binomial)

par(mfrow = c(2,3))
plot(gam.b5.rev, se=T, col="blue")
preds = predict(gam.b5.rev, newdata=data.frame(education="5. Advanced Degree", maritl="2. Married", jobclass="2. Info
rmation", age=50, health="2. >=Very Good"), type="response")
```