BUDA530Module4

## Problem 1 :

Using the wage data discussed at length in this section, we want to build a predictive model that will predict two features.

The first is we want to build a model using non-linear trends and functions to predict wage of an individual. All variables in the ISLR data are available to you and any method you know.

Subset the data such that wages below dollar 250k are the only observations in the data set. Also using these trends look at the prediction accuracy of models developed to predict whether someone’s salary is above or below (logistic regression), the median ($104k) wage in this data.

library(ISLR)  
library(gam)

## Warning: package 'gam' was built under R version 3.6.2

## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.16.1

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#head(Wage)  
#str(Wage)  
#summary(Wage)  
  
Wage250less <- filter(Wage,wage < 250)  
  
model1<-lm(wage~age+year+maritl+race+education+jobclass,data=Wage250less)  
model2 <- lm(wage~age+year+maritl+education+jobclass,data=Wage250less)  
summary(model2)

##   
## Call:  
## lm(formula = wage ~ age + year + maritl + education + jobclass,   
## data = Wage250less)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -93.208 -16.681 -1.159 15.618 98.642   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.207e+03 4.855e+02 -4.546 5.70e-06 \*\*\*  
## age 2.369e-01 4.772e-02 4.966 7.24e-07 \*\*\*  
## year 1.131e+00 2.421e-01 4.671 3.13e-06 \*\*\*  
## maritl2. Married 1.704e+01 1.333e+00 12.786 < 2e-16 \*\*\*  
## maritl3. Widowed 5.270e+00 6.220e+00 0.847 0.39690   
## maritl4. Divorced 6.386e+00 2.250e+00 2.838 0.00457 \*\*   
## maritl5. Separated 1.069e+01 3.796e+00 2.816 0.00490 \*\*   
## education2. HS Grad 1.027e+01 1.830e+00 5.611 2.20e-08 \*\*\*  
## education3. Some College 2.184e+01 1.937e+00 11.279 < 2e-16 \*\*\*  
## education4. College Grad 3.367e+01 1.945e+00 17.306 < 2e-16 \*\*\*  
## education5. Advanced Degree 4.798e+01 2.172e+00 22.091 < 2e-16 \*\*\*  
## jobclass2. Information 2.949e+00 1.028e+00 2.868 0.00416 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.43 on 2909 degrees of freedom  
## Multiple R-squared: 0.3177, Adjusted R-squared: 0.3152   
## F-statistic: 123.2 on 11 and 2909 DF, p-value: < 2.2e-16

fit1 <- gam(wage ~ s(age,2) + year + maritl + education + jobclass, data=Wage250less )

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

fit2 <- gam(wage ~ s(age,2) + s(year,3) + maritl + education + jobclass, data=Wage250less )

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

fit3 <- gam(wage ~ s(age,3) + s(year,4) + maritl + education + jobclass, data=Wage250less )

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

fit4 <- gam(wage ~ s(age,4) + s(year,5) + maritl + education + jobclass, data=Wage250less )

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

fit5 <- gam(wage ~ s(age,5) + s(year,6) + maritl + education + jobclass, data=Wage250less )

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

fit6 <- gam(wage ~ s(age,6) + s(year,6) + maritl + education + jobclass, data=Wage250less )

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

anova(fit1,fit2,fit3,fit4,fit5,fit6)

## Analysis of Deviance Table  
##   
## Model 1: wage ~ s(age, 2) + year + maritl + education + jobclass  
## Model 2: wage ~ s(age, 2) + s(year, 3) + maritl + education + jobclass  
## Model 3: wage ~ s(age, 3) + s(year, 4) + maritl + education + jobclass  
## Model 4: wage ~ s(age, 4) + s(year, 5) + maritl + education + jobclass  
## Model 5: wage ~ s(age, 5) + s(year, 6) + maritl + education + jobclass  
## Model 6: wage ~ s(age, 6) + s(year, 6) + maritl + education + jobclass  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 2908 1973542   
## 2 2906 1971564 2.00017 1977.7 0.2308797   
## 3 2904 1959754 1.99975 11810.8 0.0001576 \*\*\*  
## 4 2902 1957313 2.00027 2440.6 0.1638396   
## 5 2900 1956200 1.99997 1113.4 0.4381004   
## 6 2899 1955469 0.99972 730.7 0.2978807   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(fit1,fit2,fit3,fit4,fit5,fit6)

## df AIC  
## fit1 13 27349.68  
## fit2 13 27350.75  
## fit3 13 27337.20  
## fit4 13 27337.56  
## fit5 13 27339.90  
## fit6 13 27340.81

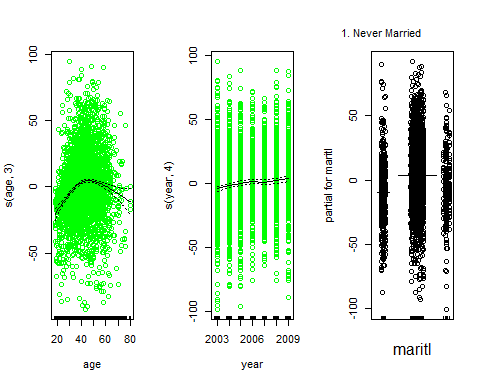
model <- gam(wage ~ s(age,3) + s(year,4) + maritl + education + jobclass, data=Wage250less)

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

summary(model)

##   
## Call: gam(formula = wage ~ s(age, 3) + s(year, 4) + maritl + education +   
## jobclass, data = Wage250less)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -96.7369 -16.2461 -0.9713 15.3071 99.1956   
##   
## (Dispersion Parameter for gaussian family taken to be 674.8463)  
##   
## Null Deviance: 2978963 on 2920 degrees of freedom  
## Residual Deviance: 1959754 on 2904 degrees of freedom  
## AIC: 27337.2   
##   
## Number of Local Scoring Iterations: 2   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## s(age, 3) 1 129476 129476 191.8604 < 2.2e-16 \*\*\*  
## s(year, 4) 1 16716 16716 24.7701 6.836e-07 \*\*\*  
## maritl 4 111001 27750 41.1207 < 2.2e-16 \*\*\*  
## education 4 607184 151796 224.9341 < 2.2e-16 \*\*\*  
## jobclass 1 4998 4998 7.4062 0.006539 \*\*   
## Residuals 2904 1959754 675   
## ---  
## 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, 3) 2 51.754 <2e-16 \*\*\*  
## s(year, 4) 3 1.258 0.2871   
## maritl   
## education   
## jobclass   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(1,3))  
plot(model,se = TRUE, col = "green", residuals = TRUE)



plot(wage~age,data=Wage250less)  
newdata1 <- data.frame(age=35, year=2006, maritl=c("1. Never Married", "2. Married", "3. Widowed", "4. Divorced", "5. Separated"), education="2. HS Grad", jobclass="1. Industrial", type="Prob")  
newdata0<- data.frame(age=18:80, year=2006, maritl="1. Never Married", education="2. HS Grad", jobclass="1. Industrial")  
#levels(Wage$jobclass)  
p1=predict(model,newdata=newdata0)  
lines(18:80,p1, col = "red", lwd=4)  
newdata2 <- data.frame(age=60, year=2009, maritl=c("1. Never Married", "2. Married", "3. Widowed", "4. Divorced", "5. Separated"), education="3. Some College", jobclass="2. Information", type="Prob")  
predict(model,newdata=newdata2)

## 1 2 3 4 5   
## 103.4332 116.5874 106.9399 105.8857 109.7825

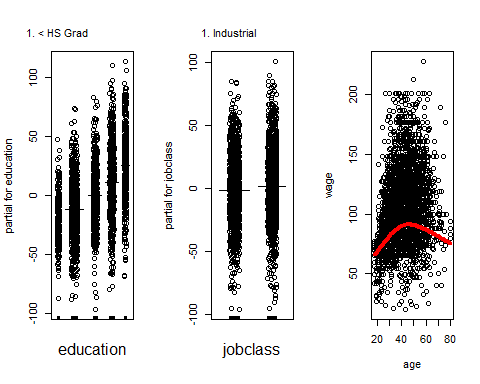
#model\_logistic <- gam(Wage104Less ~ s(age,3) + s(year,4) + maritl + education + jobclass, data=Wage250less,family=binomial)  
#summary(model\_logistic)  
#par(mfrow=c(1,1))  
#plot(model\_logistic, se = TRUE, col = "green", residuals = TRUE)  
#newdata3 <- data.frame(wage=110, age=35, year=2009, maritl=c("1. Never Married", "2. Married", "3. Widowed", "4. Divorced", "5. Separated"), education="2. HS Grad", jobclass="1. Industrial", type="Prob")  
#newdata4 <- data.frame(age=45, year=2009, maritl=("1. Never Married"), education="5. Advanced Degree", jobclass=c("1. Industrial", "2. Information"), type="Prob")  
#predict(model\_logistic, newdata =newdata3, type="response")  
  
#library(faraway)  
#ilogit(predict(model\_logistic, newdata =newdata3))  
#sum(residuals(model\_logistic,type="pearson")^2)/(df.residual(model\_logistic)+1)  
  
library(caret)

## Warning: package 'caret' was built under R version 3.6.2

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.2



set.seed(4321)  
Wage250less$Wage104Less <- ifelse(Wage250less$wage > 104, 0, 1)  
table(Wage250less$Wage104Less)

##   
## 0 1   
## 1476 1445

#1476/2921  
trainIndex <- createDataPartition(Wage250less$wage, p=.75, list=FALSE, times=1)  
Train <- Wage250less[trainIndex,]  
nrow(Train)

## [1] 2193

Test <- Wage250less[-trainIndex,]  
nrow(Test)

## [1] 728

modelTrain <- gam(Wage104Less ~ s(age,3) + s(year,4) + maritl + education + jobclass, data=Train,family=binomial)

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

summary(modelTrain)

##   
## Call: gam(formula = Wage104Less ~ s(age, 3) + s(year, 4) + maritl +   
## education + jobclass, family = binomial, data = Train)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -2.3914 -0.8859 -0.4252 0.9313 2.3170   
##   
## (Dispersion Parameter for binomial family taken to be 1)  
##   
## Null Deviance: 3039.519 on 2192 degrees of freedom  
## Residual Deviance: 2374.687 on 2176 degrees of freedom  
## AIC: 2408.687   
##   
## Number of Local Scoring Iterations: 6   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## s(age, 3) 1 25.40 25.398 25.7420 4.233e-07 \*\*\*  
## s(year, 4) 1 3.84 3.838 3.8898 0.04871 \*   
## maritl 4 34.59 8.648 8.7651 5.114e-07 \*\*\*  
## education 4 338.95 84.739 85.8881 < 2.2e-16 \*\*\*  
## jobclass 1 2.66 2.661 2.6973 0.10066   
## Residuals 2176 2146.88 0.987   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Anova for Nonparametric Effects  
## Npar Df Npar Chisq P(Chi)   
## (Intercept)   
## s(age, 3) 2 50.736 9.614e-12 \*\*\*  
## s(year, 4) 3 3.275 0.3511   
## maritl   
## education   
## jobclass   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

predictTest <- predict(modelTrain,newdata=Test, type="response")  
summary(predictTest)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.06816 0.25754 0.52939 0.50736 0.71988 0.98233

table\_matrix <- table(Test$Wage104Less,predictTest > 0.5)  
sum(diag(table\_matrix)) / sum(table\_matrix)

## [1] 0.6923077

sum(residuals(modelTrain,type="pearson")^2)/(df.residual(modelTrain)+1)

## [1] 0.9861642

### Interpretaion :

As per the summary of the different models, variables age, education, maritl status, education and jobclass are signification at p-value level 0.05. Hence, we reject the null hypothesis of these 5 variables. From the ANOVA and AIC results, Model 3 with age having degree of freedom = 3 and year having df = 4 is prefered. p-value of Model3 < 0.05 hence we reject the null hypothesis.

The age plot indicates that keeping all the predictors as fixed , wage tends to be highest for intermediate values of age (30-45), and lowest for the very young (<30) and old (>65).

The year plot indicates that keeping all the predictors Zero, wage increases in the early years and then decreases in 2006 and 2007 and again increases in later years.

The education plot indicates that wage tends to increase with education. In other words, the more educated a person is, the higher their salary.

The marriage status plot indicates that, married workers tend to make the most. The widowed, divorced, and separated groups make a small portion of the workers. The never married workers look like they make less than married workers.

The jobclass plot indicates that information workers tend to have higher wages than industrial workers with other variables set to zero.

Predicting the probablities of variables at different values,all five variables are making difference on the response and shows similar behaviour as described above.

Logistic Model Prediction accuracy : As per the Summary of the logistic model, varaibles age, year, maritl status and education are significant as p-values are < 0.05. JobClass is significant at level 0.1. All five variables are making difference on the binomial response Wage.

There are total 728 observations in test Set, out of which 361 of them are having wages < 104K , and 367 of them are having wages > 104K. In Actual, 50% of the workers are having wages < 104K. While predicting the probablity of test set, 49% of workers are predicted to have wage < 104K. Sigma2 is around 1, model fits the data fairly well.

The model can accurately identify workers with wages above and below 104K wage with test set accuracy being equal to 69%