Chapter 7 Problem 6

Andira Putri

In this exercise, you will further analyze the `Wage` data set considered throughout this chapter.

a. Perform polynomial regression to predict wage using age. Use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data.

```
set.seed(1)
library(ISLR)
data(Wage)

#perform polynomial regression up to degree 5
fit.1=lm(wage~age ,data=Wage) #linear
fit.2=lm(wage~poly(age,2),data=Wage) #quadratic
fit.3=lm(wage~poly(age,3),data=Wage) #cubic
fit.4=lm(wage~poly(age,4),data=Wage) #quartic
fit.5=lm(wage~poly(age,5),data=Wage) #quintic
anova(fit.1,fit.2,fit.3,fit.4,fit.5)
```

```
## 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)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.8888  0.001679 **
## 4   2995 4771604  1      6070   3.8098  0.051046 .
## 5   2994 4770322  1      1283   0.8050  0.369682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear and quadratic models are not sufficient because their p-values are close to 0. The quintic model has a rather high p-value at 0.37. The ANOVA table suggests that the cubic and quartic models are the best ones. Now, let's use cross-validation methods to pick the best among those two…

```
set.seed(1)
library(ISLR)
library(boot)
cv=rep(0,5)

for (i in 1:5) {
  fit=glm(wage~poly(age,i),data=Wage)
  cv[i]=cv.glm(Wage,fit,K=5)$delta
}
```

```
## Warning in cv[i] <- cv.glm(Wage, fit, K = 5)$delta: number of items to
## replace is not a multiple of replacement length
```

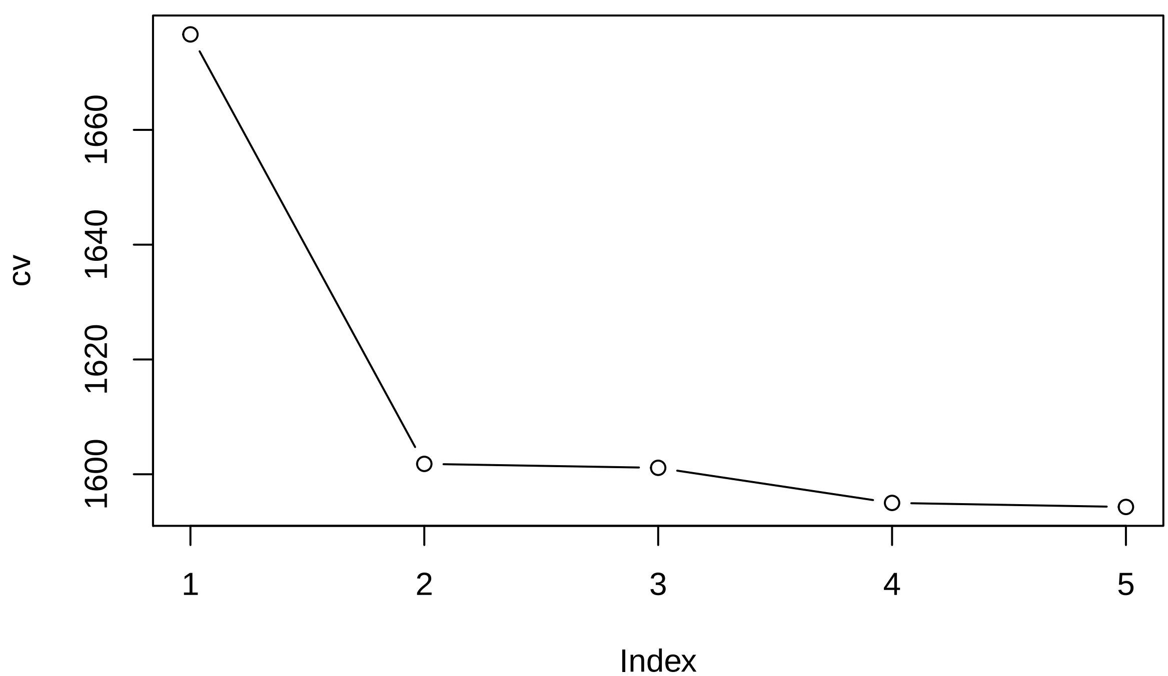
```
## Warning in cv[i] <- cv.glm(Wage, fit, K = 5)$delta: number of items to
## replace is not a multiple of replacement length

## Warning in cv[i] <- cv.glm(Wage, fit, K = 5)$delta: number of items to
## replace is not a multiple of replacement length

## Warning in cv[i] <- cv.glm(Wage, fit, K = 5)$delta: number of items to
## replace is not a multiple of replacement length

## Warning in cv[i] <- cv.glm(Wage, fit, K = 5)$delta: number of items to
## replace is not a multiple of replacement length
```

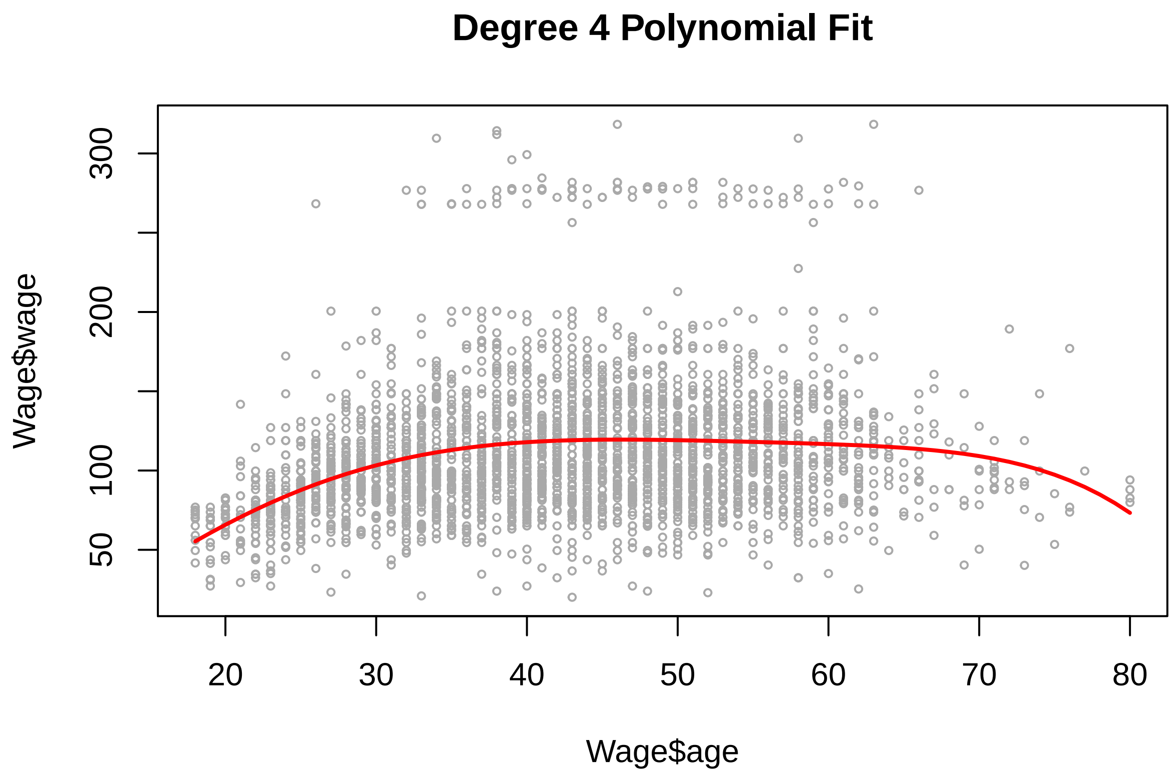
```r
plot(cv,type="b")
```



Cross-validation supports that the fourth-degree polynomial has the lowest estimated error.

From both the ANOVA table and CV, we conclude that the quartic model is the best one.

```r
#Resulting plot
library(ISLR)
agelims=range(Wage$age)
age.grid=seq(from=agelims[1],to=agelims[2])
preds=predict(fit,newdata=list(age=age.grid),se=TRUE)
plot(Wage$age,Wage$wage,xlim=agelims,cex=0.5,col="darkgrey")
title("Degree 4 Polynomial Fit")
lines(age.grid,preds$fit,lwd=2,col="red")
```

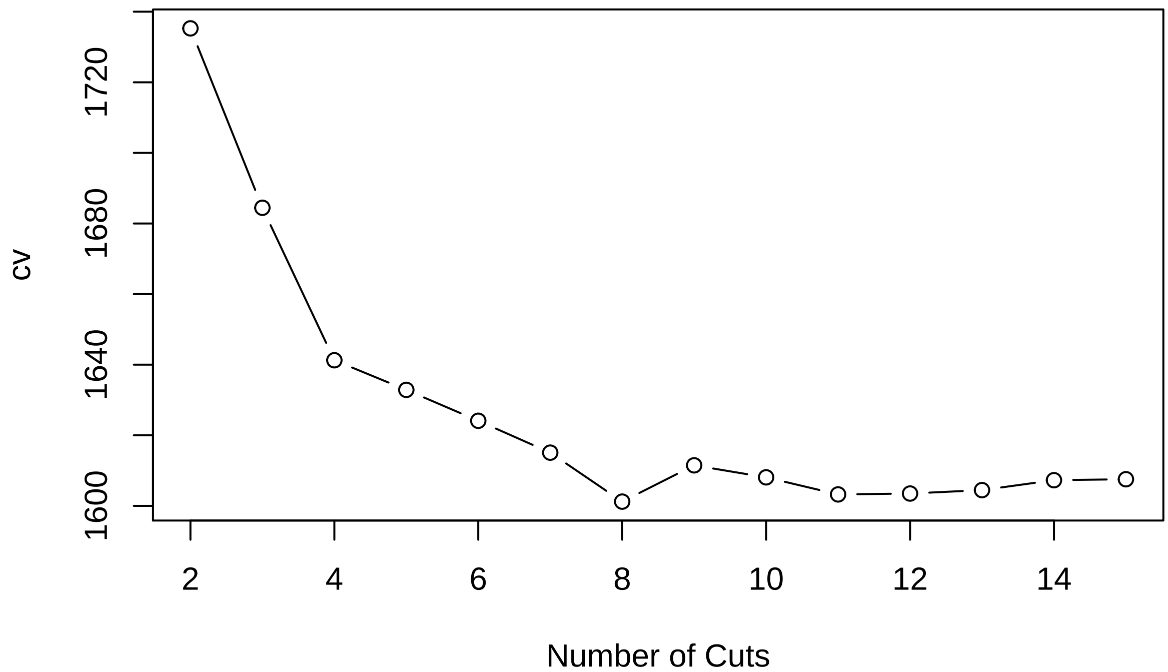


b. Fit a step function to predict wage using age, and perform cross validation to choose the optimal number of cuts. Make a plot of the fit obtained.

```
set.seed(1)
library(ISLR)
data(Wage)
cv = rep(0,9)
for (i in 2:15) {
  Wage$age.cut=cut(Wage$age,i)
  step=glm(wage~age.cut,data=Wage)
  cv[i-1]=cv.glm(Wage,step,K=5)$delta[1]
}
cv
```

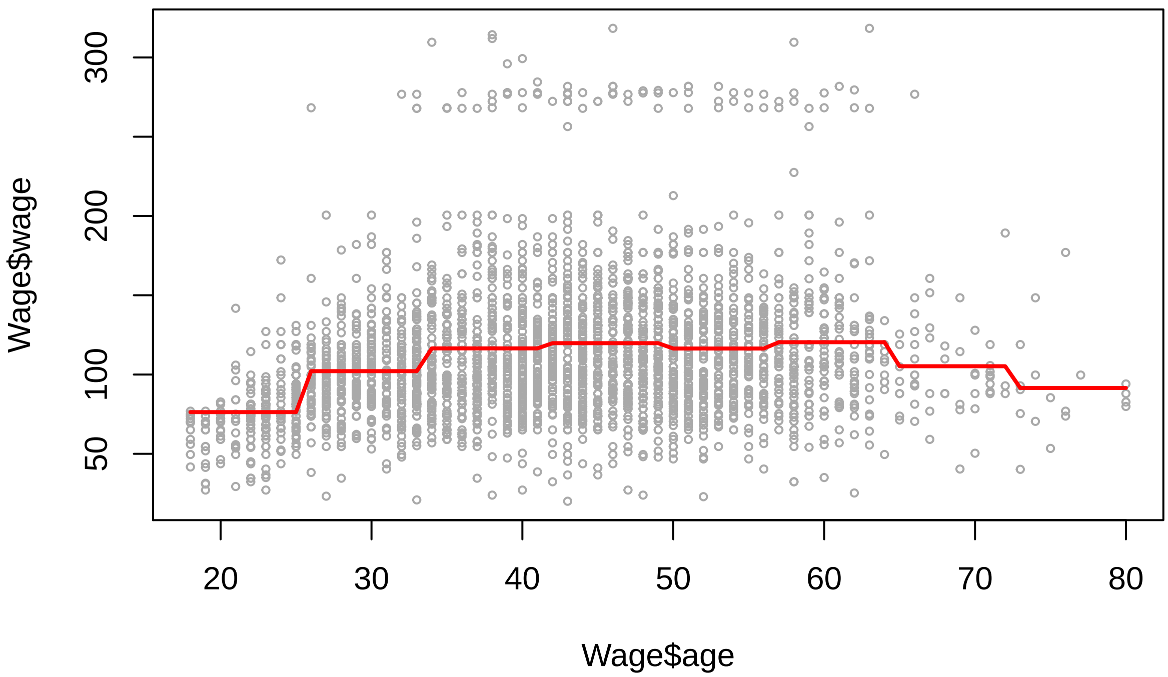
```
##  [1] 1735.276 1684.451 1641.254 1632.860 1624.106 1615.079 1601.205
##  [8] 1611.484 1608.079 1603.243 1603.490 1604.463 1607.274 1607.543
```

```
plot(2:15, cv, type="b",xlab="Number of Cuts")
```



With the plot, it seems like 8 is the best number of cuts.

```
#Step Function Plot
cut=glm(wage~cut(age,8), data=Wage)
preds=predict(cut, newdata=list(age=age.grid), se=TRUE)
plot(Wage$age, Wage$wage, xlim=agelims, cex=0.5, col="darkgrey")
lines(age.grid, preds$fit, lwd=2, col="red")
```


Wage$wage
300
200
100
50
20
30
40
50
60
70
80
Wage$age