Q28-WilliamKennedy-300015367

William Kennedy

2023-11-10

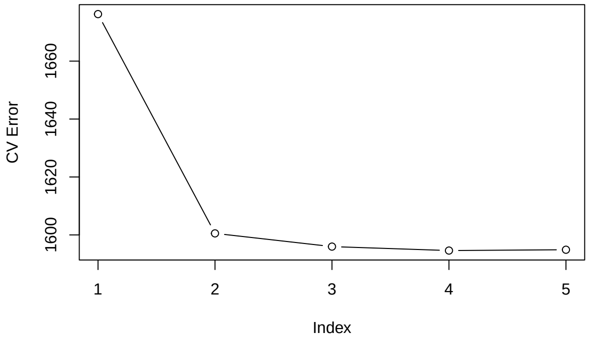
```
Wage = read.csv("Wage.csv")
wage = Wage$wage
age = Wage$age

wage.lm = lm(wage~age,data=Wage)
```

1. 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.

```
library("boot")
```

```
## Warning: package 'boot' was built under R version 4.3.2
```

```
 cv.error <- rep (0, 5)
for (i in 1:5) {
  glm.fit <- glm (wage~poly(age , i), data = Wage)
  cv.error[i] <- cv.glm (Wage , glm.fit)$delta[1]
}
plot(cv.error,type='b',ylab="CV Error")
```

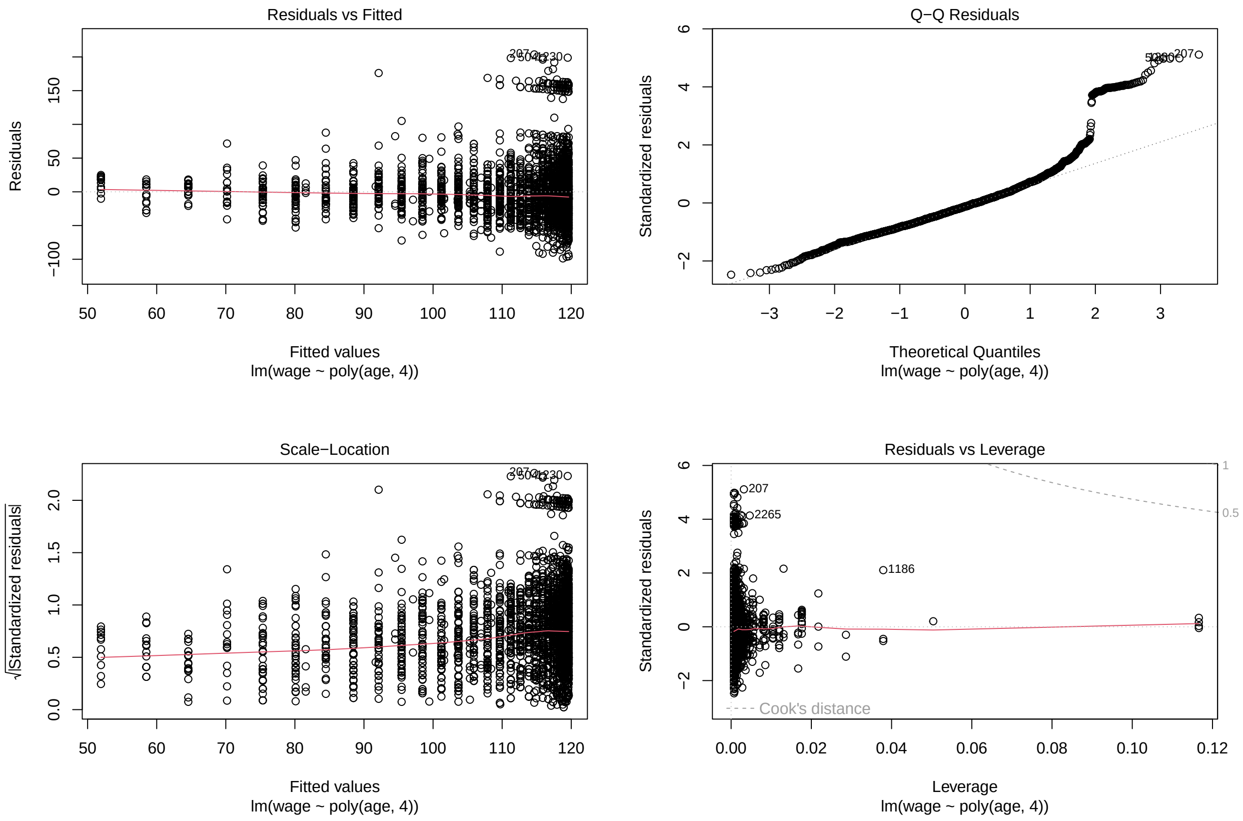


```
print(cv.error)
```

```
## [1] 1676.235 1600.529 1595.960 1594.596 1594.879
```

The degree with the smallest cross validation error is d=4, which has a cross validation error of 1594.596

```
fit.1 <- lm(wage~age , data = Wage)
fit.2 <- lm(wage~poly (age , 2), data = Wage)
fit.3 <- lm(wage~poly (age , 3), data = Wage)
fit.4 <- lm(wage~poly (age , 4), data = Wage)
fit.5 <- lm(wage~poly (age , 5), data = Wage)
anova= anova (fit.1, fit.2, fit.3, fit.4, fit.5)

plot(fit.4)
```

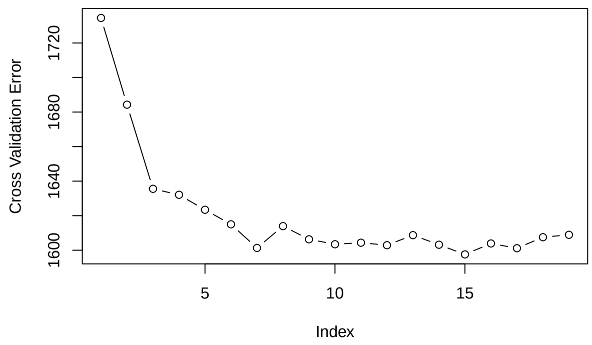


2. Fit a step function to predict wage using age, and perform cross-validation to choose the optimal number of cuts. Make a plot of the fit obtained.

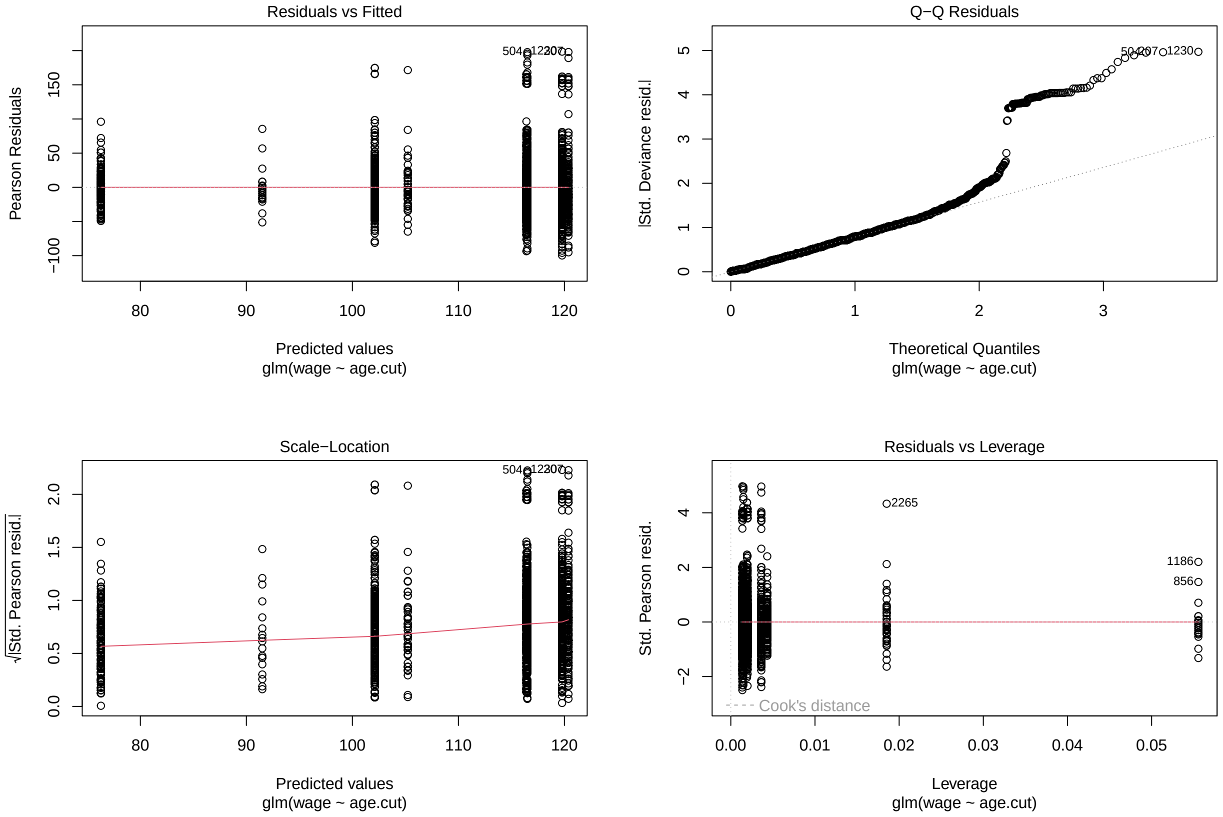
```
library('ggplot2')
set.seed(1)
cv.error.cut = rep(NA,19)

for (i in 2:20) {
  Wage$age.cut = cut(Wage$age,i)
  step.fit=glm(wage~age.cut,data=Wage)
  cv.error.cut[i-1]=cv.glm(Wage,step.fit,K=10)$delta[1] # [1]: Std [2]: Bias corrected.
}
cv.error.cut
```

```
##  [1] 1734.489 1684.271 1635.552 1632.080 1623.415 1614.996 1601.318 1613.954
##  [9] 1606.331 1603.465 1604.349 1602.915 1608.731 1603.178 1597.583 1603.909
## [17] 1601.161 1607.540 1608.915
```

```
plot(cv.error.cut,type='b',ylab="Cross Validation Error")
```



```
Wage$age.cut = cut(Wage$age, 8)
fit=glm(wage~age.cut,data=Wage)
plot(fit)
```



Hence the optimal number of cuts is 8, since the cross validation error doesn't improve after 8 cuts.