Assignment28

John Bute

2024-11-08

R Markdown

```
Wage <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/Wage.csv")
```

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

1. Perform polynomial regression to predict wage using age. Use cross-validation to select the optimaldegree 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)
library(ggplot2)
library(splines)
```

Polynomial regression:

```
set.seed(1)

cv_error <- sapply(1:10, function(d){
  glm_fit <- glm(wage ~ poly(age, d), data = Wage)
  cv.glm(Wage, glm_fit, K =10)$delta[1]
})

optimal_degree <- which.min(cv_error)
optimal_degree

## [1] 9
```

We found the optimal degree of 9.

Hypothesis testing using ANOVA:

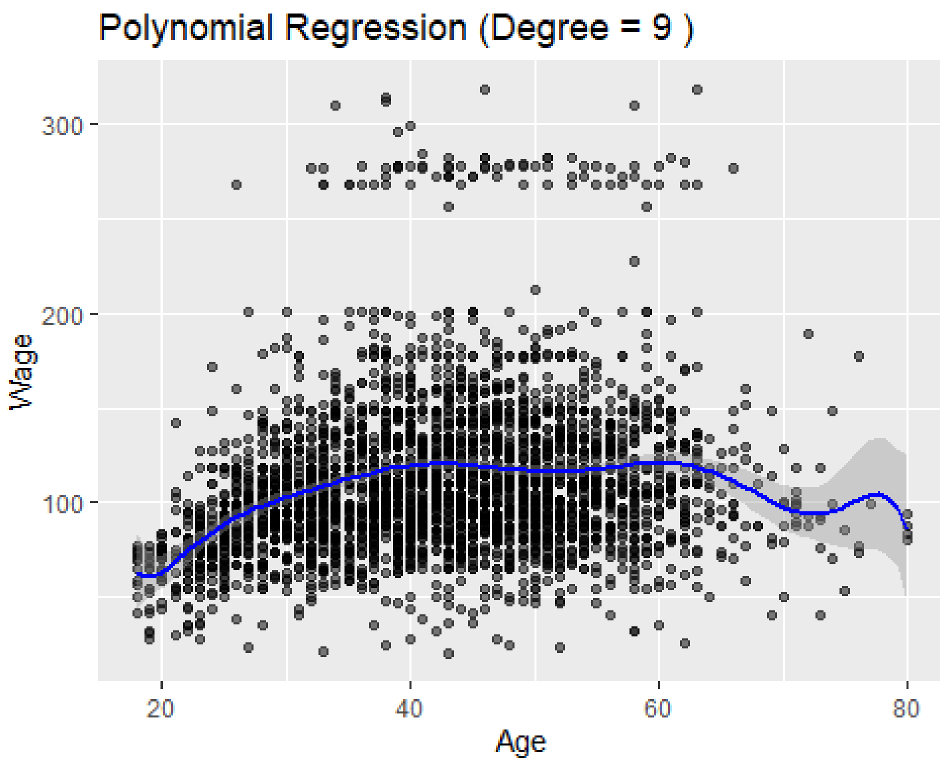
```
library(ggplot2)



fit1 <- lm(wage ~ poly(age, 1), data = Wage)
fit2 <- lm(wage ~ poly(age, 2), data = Wage)
fit3 <- lm(wage ~ poly(age, 3), data = Wage)
fit4 <- lm(wage ~ poly(age, 4), data = Wage)
fit5 <- lm(wage ~ poly(age, 5), data = Wage)
fit6 <- lm(wage ~ poly(age, 6), data = Wage)
fit7 <- lm(wage ~ poly(age, 7), data = Wage)
```

```r
fit8 <- lm(wage ~ poly(age, 8), data = Wage)
fit9 <- lm(wage ~ poly(age, 9), data = Wage)

anova_results <- anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9)
anova_results
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ poly(age, 1)
## Model 2: wage ~ poly(age, 2)
## Model 3: wage ~ poly(age, 3)
## Model 4: wage ~ poly(age, 4)
## Model 5: wage ~ poly(age, 5)
## Model 6: wage ~ poly(age, 6)
## Model 7: wage ~ poly(age, 7)
## Model 8: wage ~ poly(age, 8)
## Model 9: wage ~ poly(age, 9)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.8118 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.9038  0.001666 **
## 4   2995 4771604  1      6070   3.8156  0.050870 .
## 5   2994 4770322  1      1283   0.8062  0.369318
## 6   2993 4766389  1      3932   2.4718  0.116014
## 7   2992 4763834  1      2555   1.6062  0.205123
## 8   2991 4763707  1       127   0.0796  0.777829
## 9   2990 4756703  1      7004   4.4028  0.035963 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we increase the degrees, we see a decrease in RSS, and decrease the F-statistic. However, after degree 3, we see that the terms become statistically insignificant till we reach the degree 9, where we see a slight but significant imrpovment, which might explain while cv selected 9 as our optimal degrees of freedom. However, we would like to note that a polynomial of degree 9 might be prone to overfitting especially to noise, and this may indicate better performance on the training set but not necessarily on the test set.

```r
ggplot(Wage, aes(x = age, y = wage)) +
  geom_point(alpha = 0.5) +
  stat_smooth(method = "lm", formula = y ~ poly(x, optimal_degree), color =
"blue") +
  labs(title = paste("Polynomial Regression (Degree =", optimal_degree, ")"),
       x = "Age", y = "Wage")
```

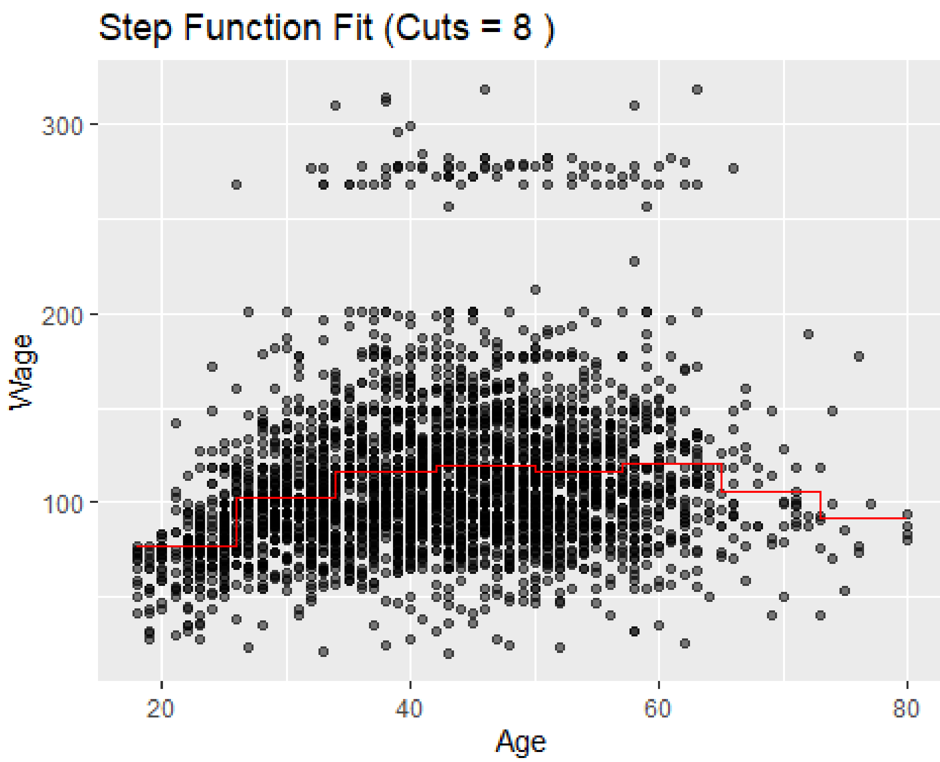


2. Fit a step function to predict wage using age, and perform cross-validation to choose the optimal number of cuts. Make a plot of the fit obtained

```
set.seed(2)

cv_step_function <- function(cuts) {
  Wage$age.cut <- cut(Wage$age, cuts)
  glm_fit <- glm(wage ~ age.cut, data = Wage)
  return(cv.glm(Wage, glm_fit, K = 10)$delta[1])
}
cv_errors_step <- sapply(2:10, cv_step_function)

optimal_cuts <- which.min(cv_errors_step) + 1
optimal_cuts

## [1] 8
```

We find that the optimal number of cuts is 8.

```
Wage$age.cut <- cut(Wage$age, optimal_cuts)

ggplot(Wage, aes(x = age, y = wage)) +
  geom_point(alpha = 0.5) +
  geom_step(aes(x = age, y = fitted(lm(wage ~ age.cut, data = Wage))), color
= "red") +
  labs(title = paste("Step Function Fit (Cuts =", optimal_cuts, ")"),
       x = "Age", y = "Wage")
```


Step Function Fit (Cuts = 8)
Wage
300
200
100
20
40
60
80
Age