Homework 6

Anthony Anderson

April 13, 2019

Problem 1

a)

```
library(ISLR)
library(boot)
set.seed(42)
degree <- 10
cv.poly.errors <- rep(NA, degree)
for (i in 1:degree) {
  poly.fit <- glm(wage ~ poly(age, i), data = Wage)
  cv.poly.errors[i] <- cv.glm(Wage, poly.fit)$delta[1]
}
cv.poly.errors
```

```
##  [1] 1676.235 1600.529 1595.960 1594.596 1594.879 1594.119 1594.145
##  [8] 1594.975 1593.356 1594.232
```

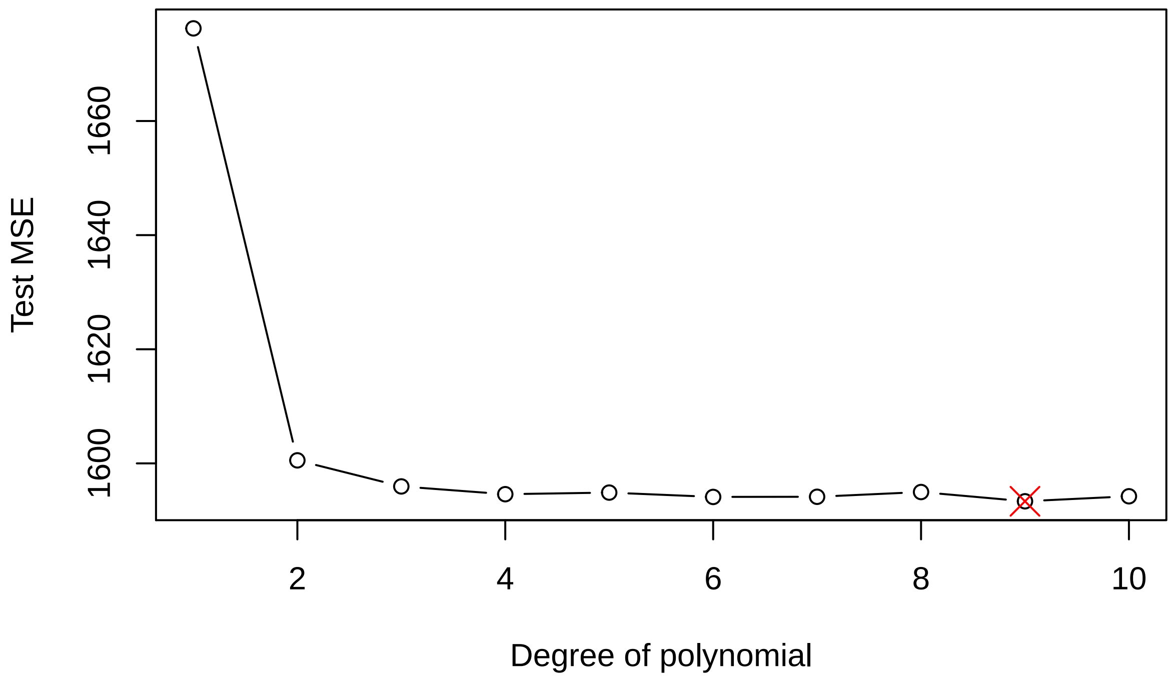
```
plot(1:degree, cv.poly.errors, xlab = "Degree of polynomial", ylab = "Test MSE", type="b")
deg.min <- which.min(cv.poly.errors)
points(deg.min, cv.poly.errors[deg.min], col = "red", cex = 2, pch = 4)

fit1 <- lm(wage ~ age, 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)
fit8 <- lm(wage ~ poly(age, 8), data = Wage)
anova(fit1,fit2,fit3,fit4,fit8)
```

```
## 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, 8)
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.6484 < 2.2e-16 ***
## 3   2996 4777674  1     15756   9.8926  0.001676 **
## 4   2995 4771604  1      6070   3.8113  0.051002 .
## 5   2991 4763707  4      7897   1.2395  0.291872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

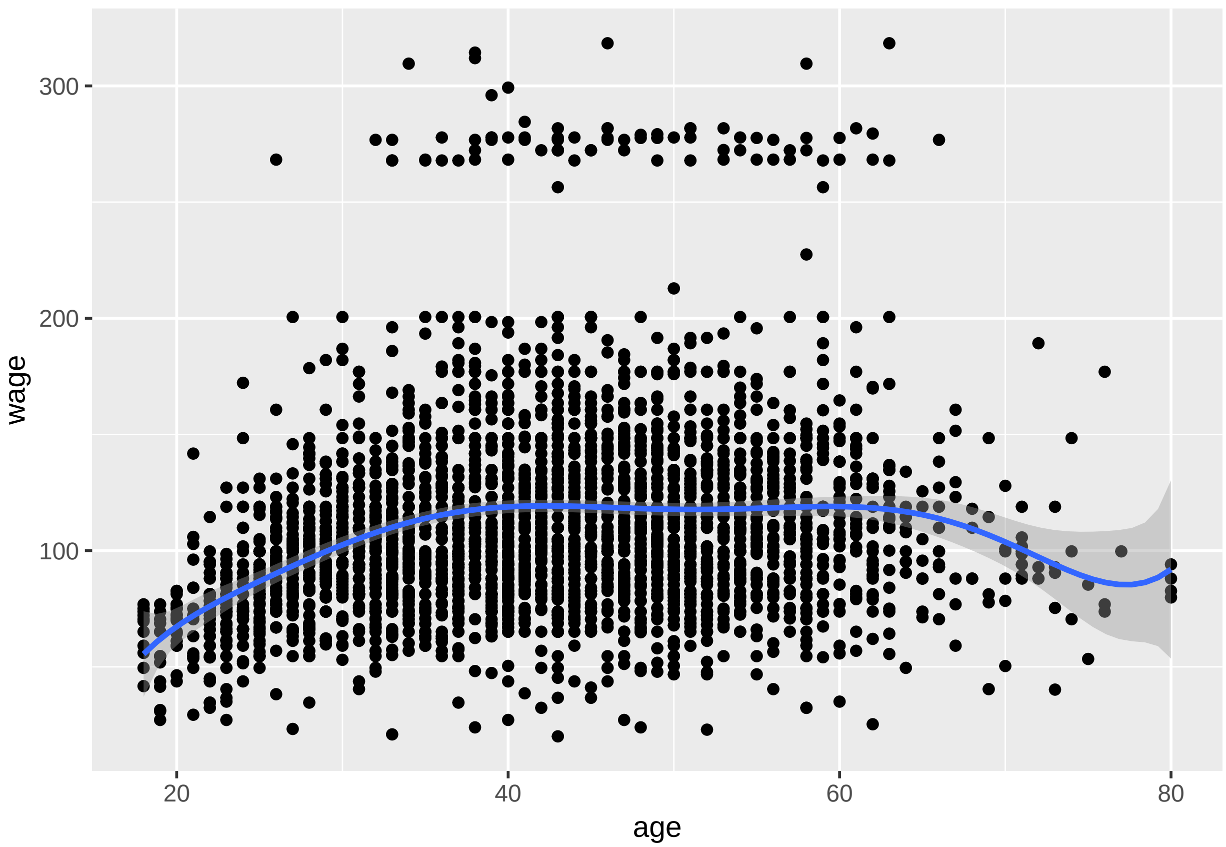
```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages --------------------------------------------------------------------- tidyvers
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'tidyr' was built under R version 3.5.3

## Warning: package 'readr' was built under R version 3.5.3

## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## -- Conflicts ------------------------------------------------------------------------ tidyverse_confl
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



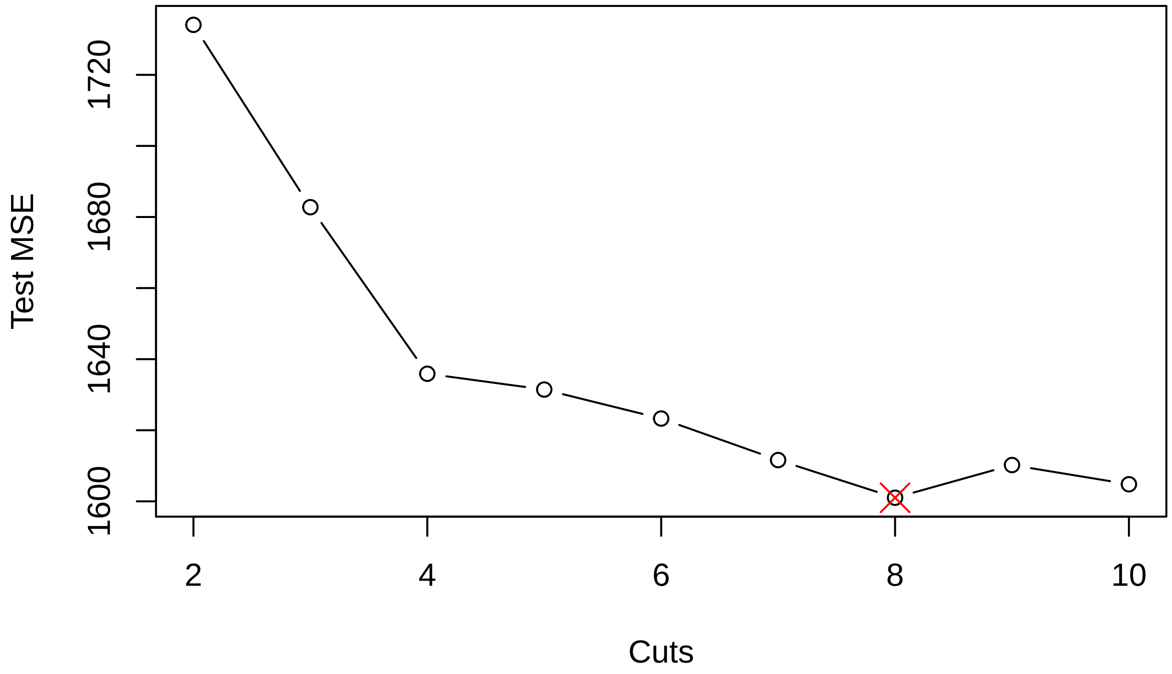
```
Wage %>% ggplot(aes(x = age, y = wage)) + geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2) + I(x^3) + I(x^4)+I(x^5) + I(x^6) + I(x^7)+I(x^8)
```



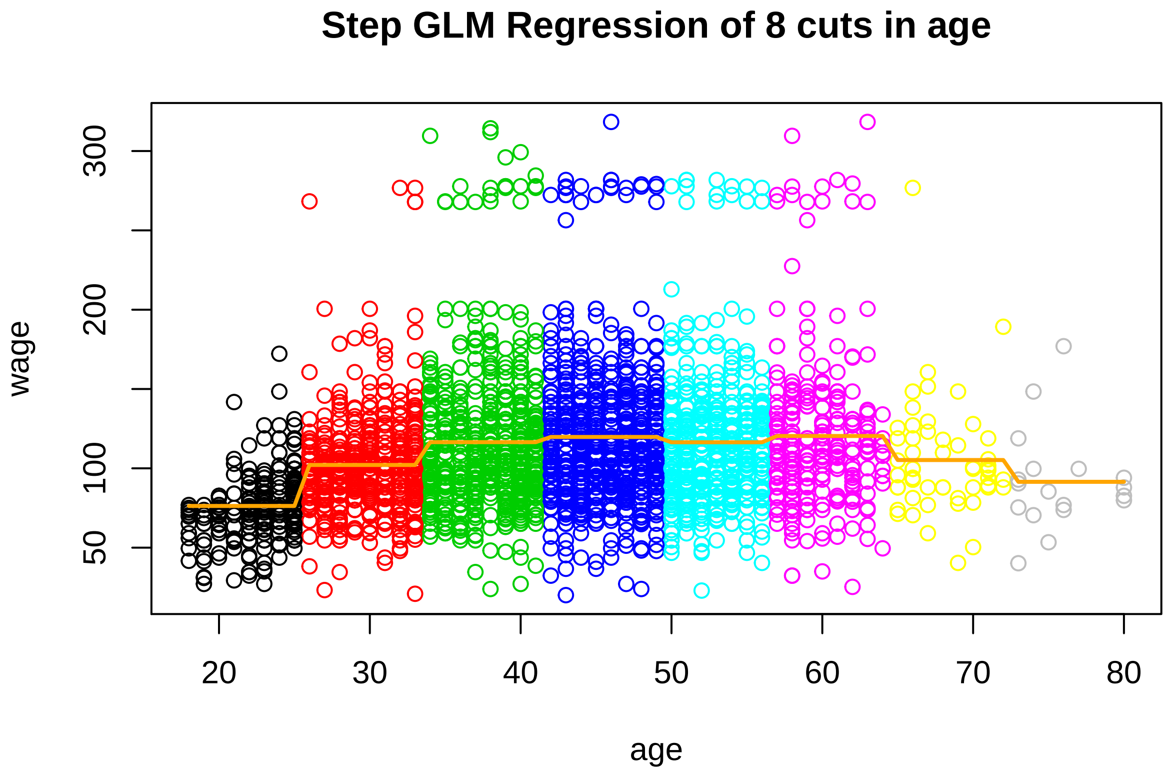
CV with K=10 suggests polynomial with degree 9 is the best since it has the lowest test MSE. However we know through principle of parisomony and the ANOVA regression output that the polynomial of degree 3 or 4 is still viable and a better choice since it has much fewer terms. ANOVA even suggests that degree 8 polynomial is not significant (p=value=.29) compared to degree 3 or 4.

b)

```
set.seed(42)
cv.step.errors <- rep(NA, degree)
for (i in 2:degree) {
  Wage$age.cut <- cut(Wage$age, i)
  step.fit <- glm(wage ~ age.cut, data = Wage)
  cv.step.errors[i] <- cv.glm(Wage, step.fit)$delta[1]
}
plot(2:degree, cv.step.errors[-1], xlab = "Cuts", ylab = "Test MSE", type = "b")
deg.min <- which.min(cv.step.errors)
points(deg.min, cv.step.errors[deg.min], col = "red", cex = 2, pch = 4)
```



```
age.grid <- seq(min(Wage$age),max(Wage$age))
age.color <- cut(Wage$age,8)
plot(wage ~ age, data = Wage, col=age.color,main="Step GLM Regression of 8 cuts in age")
step.eight.fit <- glm(wage ~ cut(age, 8), data = Wage)
step.eight.predict <- predict(step.eight.fit, list(age = age.grid))
lines(age.grid, step.eight.predict, col = "orange", lwd = 2)
```



The TEST MSE by cut plot suggests that having 8 cuts will have the lowest test MSE. A step regression with 8 cuts is plotted above, color coded by age "group".