Assignment 6

Tina Roha

4/23/2021

Moving Beyond Linearity

Chapter 07 (page 297): Questions 6 and 10

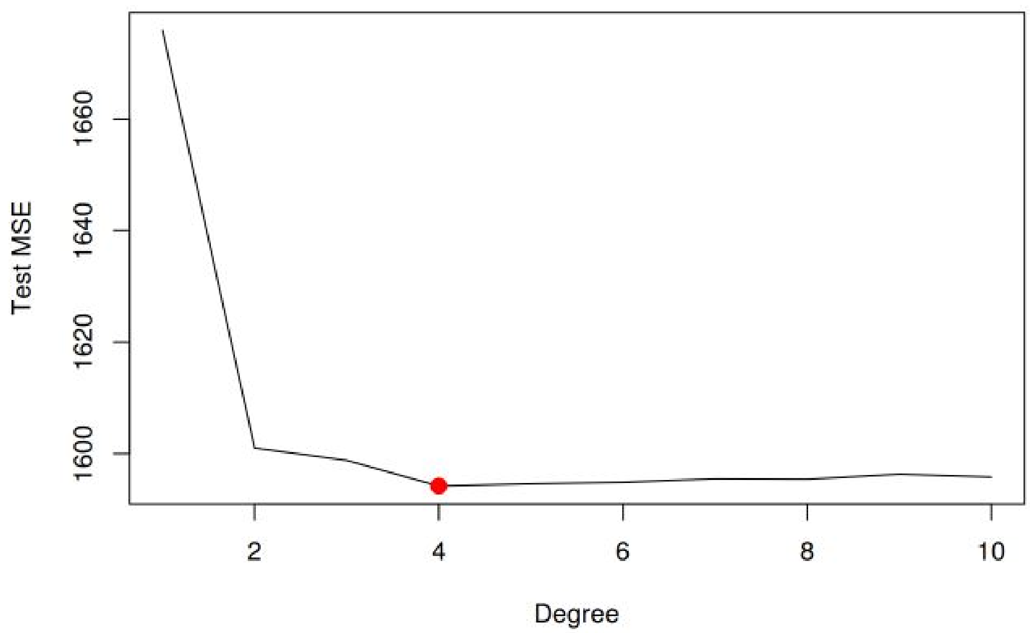
Question-6

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.

The following results will be from utilizing a K-fold cross-validation with K = 10.

```r
library(ISLR)
library(boot)
set.seed(1)
deltas <- rep(NA, 10)
for (i in 1:10) {
    fit <- glm(wage ~ poly(age, i), data = Wage)
    deltas[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(1:10, deltas, xlab = "Degree", ylab = "Test MSE", type = "l")
d.min <- which.min(deltas)
points(which.min(deltas), deltas[which.min(deltas)], col = "red", cex = 2, pch = 20)
```



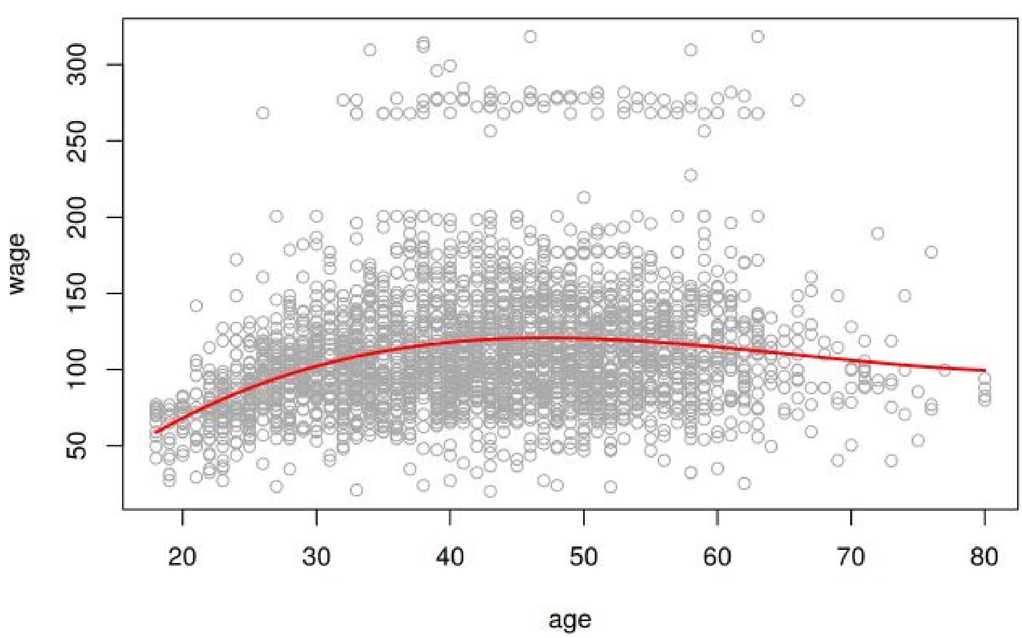
From the produced output it can be concluded that d=4 is the optimal degree.

The next step is to utilize ANOVA in order to test the null hypothesis. Furthermore, the null hypothesis states that a model M1 can sufficiently explain the data without needing a more complex M2.

```
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)
fit5 <- lm(wage ~ poly(age, 5), data = Wage)
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## 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
```

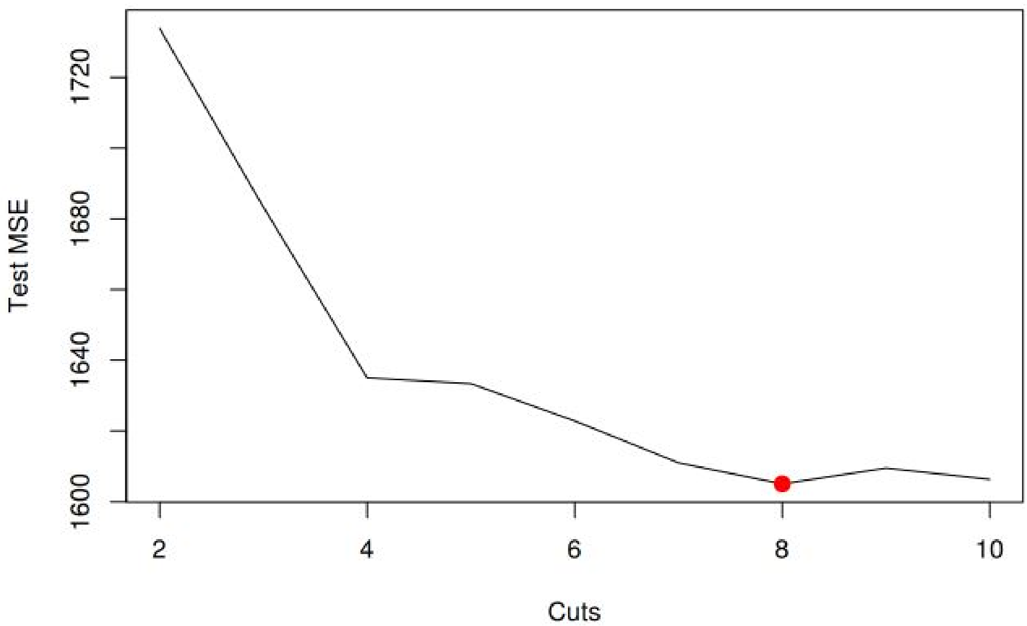
```
plot(wage ~ age, data = Wage, col = "darkgrey")
agelims <- range(Wage$age)
age.grid <- seq(from = agelims[1], to = agelims[2])
fit <- lm(wage ~ poly(age, 3), data = Wage)
preds <- predict(fit, newdata = list(age = age.grid))
lines(age.grid, preds, col = "red", lwd = 2)
```



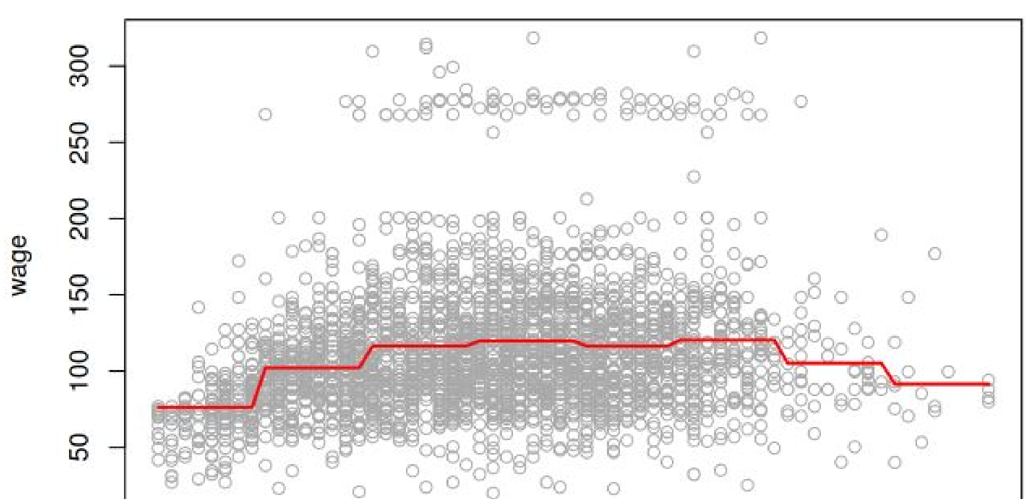
b. Fit a step function to predict wage using age, and perform crossvalidation to choose the optimal number of cuts. Make a plot of the fit obtained.

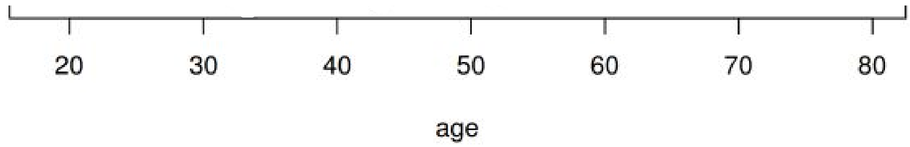
The following results will be from utilizing a K-fold cross-validation with K = 10.

```
cvs <- rep(NA, 10)
for (i in 2:10) {
    Wage$age.cut <- cut(Wage$age, i)
    fit <- glm(wage ~ age.cut, data = Wage)
    cvs[i] <- cv.glm(Wage, fit, K = 10)$delta[1]
}
plot(2:10, cvs[-1], xlab = "Cuts", ylab = "Test MSE", type = "l")
d.min <- which.min(cvs)
points(which.min(cvs), cvs[which.min(cvs)], col = "red", cex = 2, pch = 20)
```




```
plot(wage ~ age, data = Wage, col = "darkgrey")
agelims <- range(Wage$age)
age.grid <- seq(from = agelims[1], to = agelims[2])
fit <- glm(wage ~ cut(age, 8), data = Wage)
preds <- predict(fit, data.frame(age = age.grid))
lines(age.grid, preds, col = "red", lwd = 2)
```



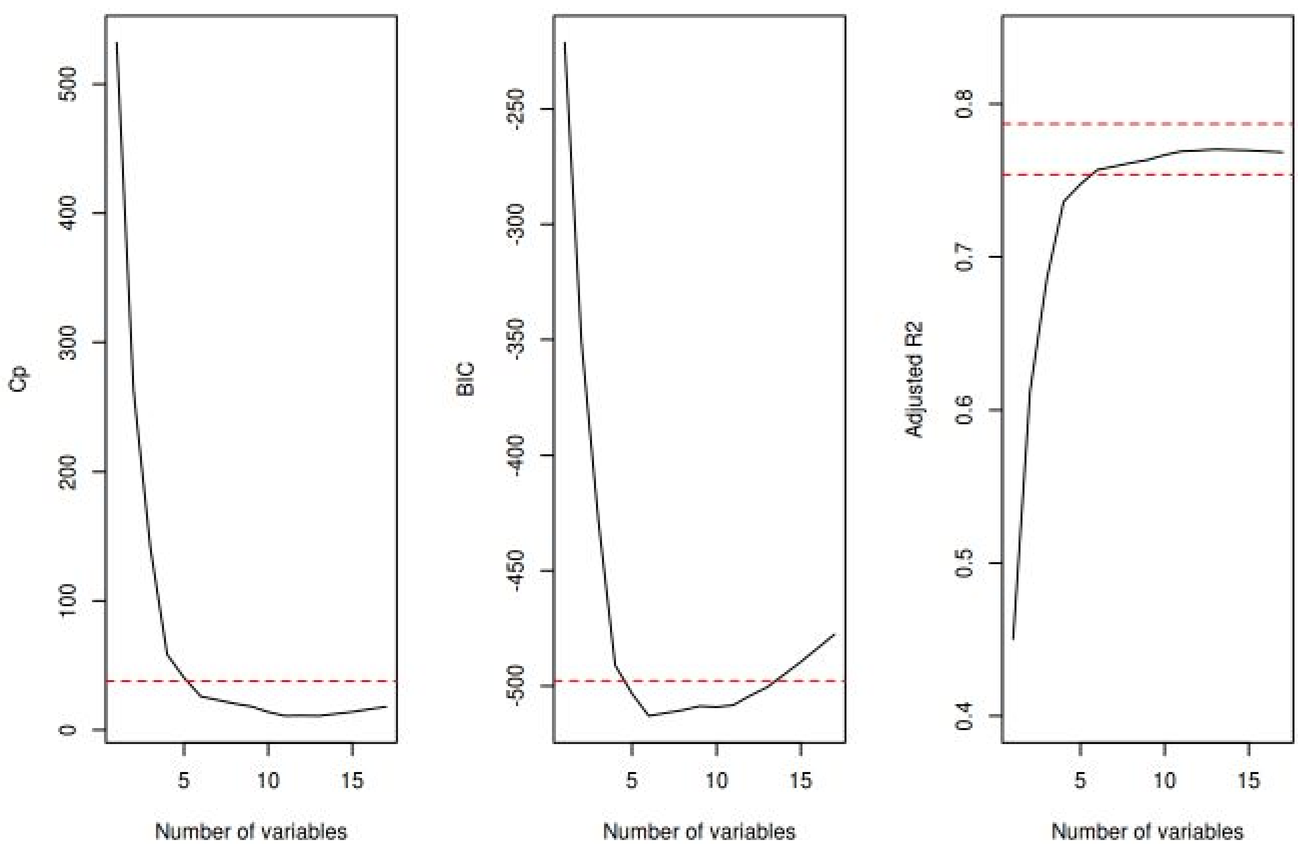


Question-10

This question relates to the College data set.

a. Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

```r
library(leaps)
set.seed(1)
attach(College)
train <- sample(length(Outstate), length(Outstate) / 2)
test <- -train
College.train <- College[train, ]
College.test <- College[test, ]
fit <- regsubsets(Outstate ~ ., data = College.train, nvmax = 17, method = "forward")
fit.summary <- summary(fit)
par(mfrow = c(1, 3))
plot(fit.summary$cp, xlab = "Number of variables", ylab = "Cp", type = "l")
min.cp <- min(fit.summary$cp)
std.cp <- sd(fit.summary$cp)
abline(h = min.cp + 0.2 * std.cp, col = "red", lty = 2)
abline(h = min.cp - 0.2 * std.cp, col = "red", lty = 2)
plot(fit.summary$bic, xlab = "Number of variables", ylab = "BIC", type='l')
min.bic <- min(fit.summary$bic)
std.bic <- sd(fit.summary$bic)
abline(h = min.bic + 0.2 * std.bic, col = "red", lty = 2)
abline(h = min.bic - 0.2 * std.bic, col = "red", lty = 2)
plot(fit.summary$adjr2, xlab = "Number of variables", ylab = "Adjusted R2", type = "l", ylim = c(0.4, 0.84))
max.adjr2 <- max(fit.summary$adjr2)
std.adjr2 <- sd(fit.summary$adjr2)
abline(h = max.adjr2 + 0.2 * std.adjr2, col = "red", lty = 2)
abline(h = max.adjr2 - 0.2 * std.adjr2, col = "red", lty = 2)
```

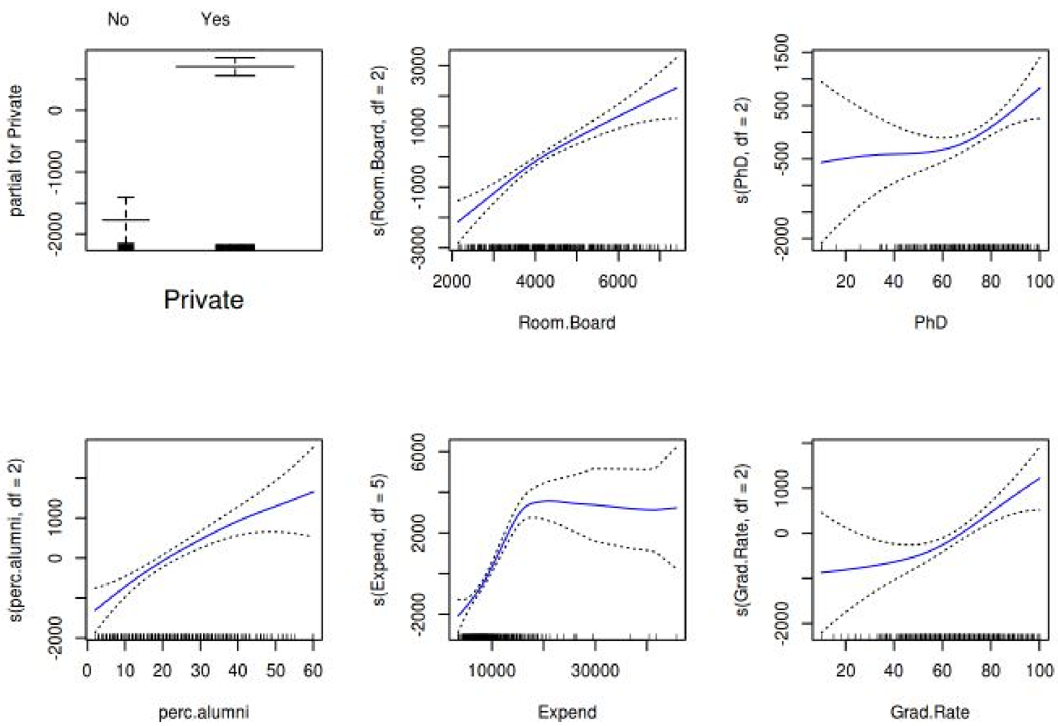



```r
fit <- regsubsets(Outstate ~ ., data = College, method = "forward")
coeffs <- coef(fit, id = 6)
names(coeffs)
```

```
## [1] "(Intercept)" "PrivateYes"  "Room.Board"  "PhD"         "perc.alumni"
## [6] "Expend"      "Grad.Rate"
```

b. Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

```
fit <- gam(Outstate ~ Private + s(Room.Board, df = 2) + s(PhD, df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5)
+ s(Grad.Rate, df = 2), data=College.train)
par(mfrow = c(2, 3))
plot(fit, se = T, col = "blue")
```



c. Evaluate the model obtained on the test set, and explain the results obtained.

```
preds <- predict(fit, College.test)
err <- mean((College.test$Outstate - preds)^2)
err
```

```
## [1] 3745460
```

```
tss <- mean((College.test$Outstate - mean(College.test$Outstate))^2)
rss <- 1 - err / tss
rss
```

```
## [1] 0.7696916
```

From the results, it can be concluded that the r-squared value is 0.77 using GAM with six predictors.

d. For which variables, if any, is there evidence of a non-linear relationship with the response?

```r
summary(fit)
```

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board, df = 2) + s(PhD,
##     df = 2) + s(perc.alumni, df = 2) + s(Expend, df = 5) + s(Grad.Rate,
##     df = 2), data = College.train)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4977.74 -1184.52    58.33  1220.04  7688.30
##
## (Dispersion Parameter for gaussian family taken to be 3300711)
##
##     Null Deviance: 6221998532 on 387 degrees of freedom
## Residual Deviance: 1231165118 on 373 degrees of freedom
## AIC: 6941.542
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                         Df     Sum Sq    Mean Sq F value    Pr(>F)
## Private                  1 1779433688 1779433688 539.106 < 2.2e-16 ***
## s(Room.Board, df = 2)    1 1221825562 1221825562 370.171 < 2.2e-16 ***
## s(PhD, df = 2)           1  382472137  382472137 115.876 < 2.2e-16 ***
## s(perc.alumni, df = 2)   1  328493313  328493313  99.522 < 2.2e-16 ***
## s(Expend, df = 5)        1  416585875  416585875 126.211 < 2.2e-16 ***
## s(Grad.Rate, df = 2)     1   55284580   55284580  16.749 5.232e-05 ***
## Residuals              373 1231165118    3300711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                        Npar Df  Npar F     Pr(F)
## (Intercept)
## Private
## s(Room.Board, df = 2)        1  3.5562   0.06010 .
## s(PhD, df = 2)               1  4.3421   0.03786 *
## s(perc.alumni, df = 2)       1  1.9158   0.16715
## s(Expend, df = 5)            4 16.8636 1.016e-12 ***
## s(Grad.Rate, df = 2)         1  3.7208   0.05450 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output it can be concluded that there is evidence of a non-linear relationship between Outstate and Expend. Additionally, there is also evidence of a non-linear relationship between Outstate, Grad.Rate or PhD.