Question 17: Problem 8 in Section 6.8

Julia Lee

10/12/2018

A

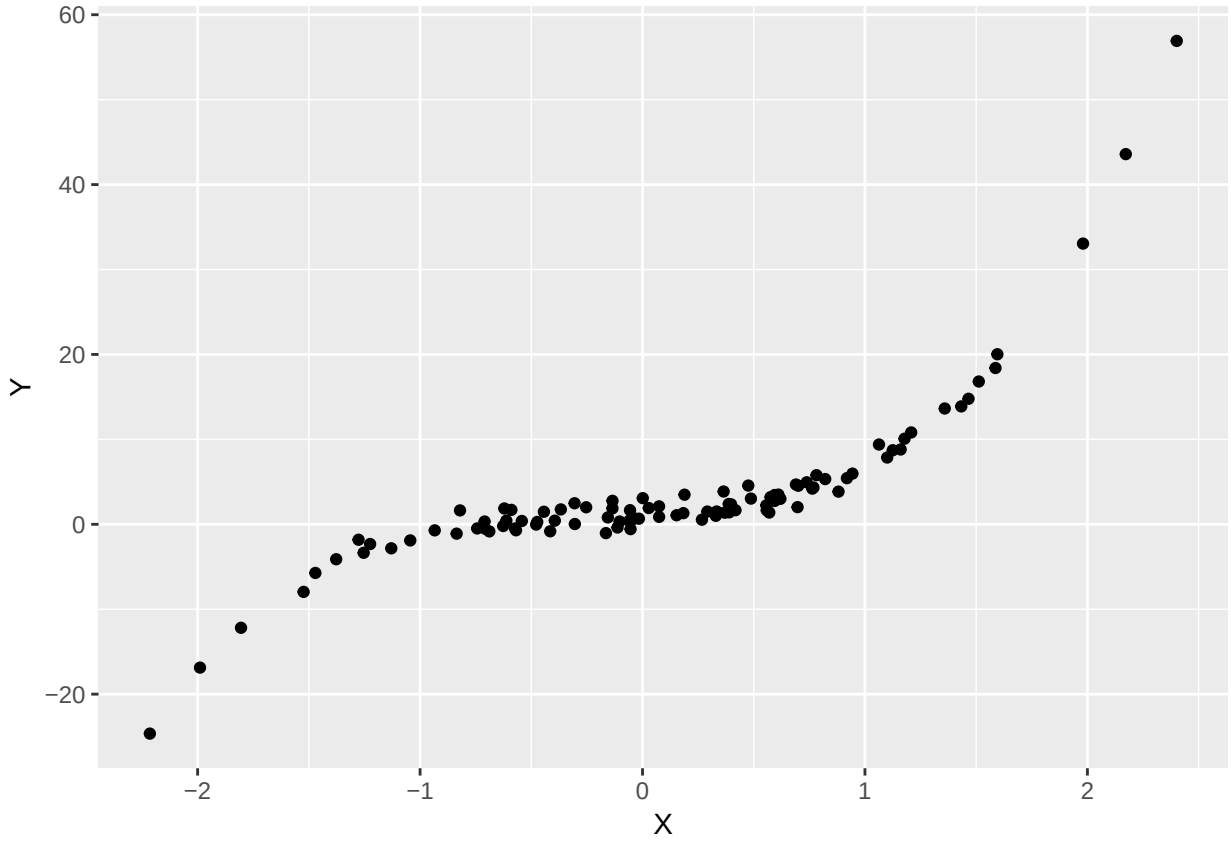
```
set.seed(1)
X <-rnorm(100)
e<-rnorm(100)
```

B

```
Y<- 1+1*X+2*X^2+3*X^3+e
```

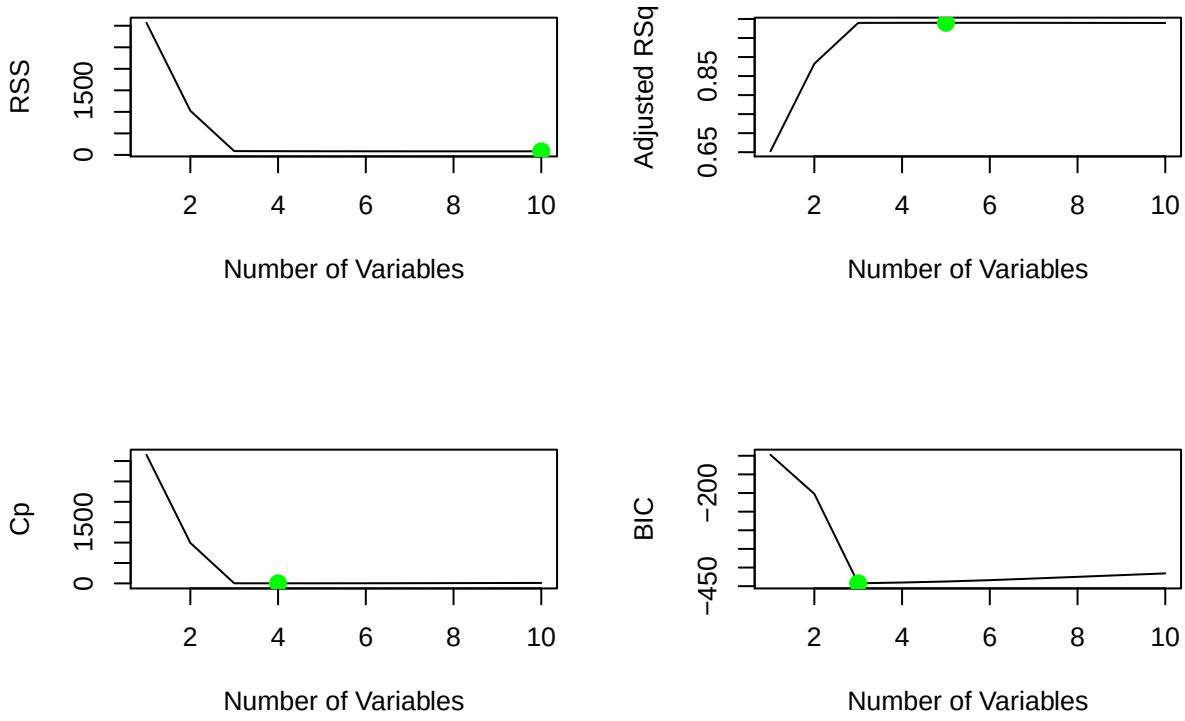
C

```
set.seed(1)
frame <- data.frame(Y, X )
best<-regsubsets(Y~poly(X,10), data = frame, nvmax = 10)
summary <- summary(best)
ggplot(data=frame, aes(x=X, y=Y))+
  geom_point()
```




```
# Set up a 2x2 grid so we can look at 4 plots at once
par(mfrow = c(2,2))
plot(summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")
rss_min <- which.min(summary$rss)
points(rss_min, summary$rss[rss_min], col = "green", cex = 2, pch = 20)

plot(summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adjr2_max <- which.max(summary$adjr2)
points(adjr2_max, summary$adjr2[adjr2_max], col = "green", cex = 2, pch = 20)
plot(summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
cp_min <- which.min(summary$cp)
points(cp_min, summary$cp[cp_min], col = "green", cex = 2, pch = 20)
plot(summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min <- which.min(summary$bic)
points(bic_min, summary$bic[bic_min], col = "green", cex = 2, pch = 20)
```




```
coef(best, which.max(summary$adjr2))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)4
##     3.324088    76.459510    30.621314    45.183672     1.257095
## poly(X, 10)5
##     1.480188
```

```
coef(best, which.min(summary$bic))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3
##     3.324088    76.459510    30.621314    45.183672
```

```
coef(best, which.min(summary$cp))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)5
##     3.324088    76.459510    30.621314    45.183672     1.480188
```

We see that **best subset selection**, using Cp we will choose the 4-variables model. The best model obtained according to Cp is the model that uses X^1, X^2, X^3, and X^5 as predictors. And using BIC we pick the model that uses three variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3 as predictors. Adjusted R^2 tells us to pick model that has 5-variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3, X^4, and X^5 as predictors. RSS tells us to choose the full ten variable model. After looking at the scatterplot of our data it looks to be a cubic relationship(we also know that the underlying relationship is cubic plus an error) so the best model would be the model obtained by BIC:

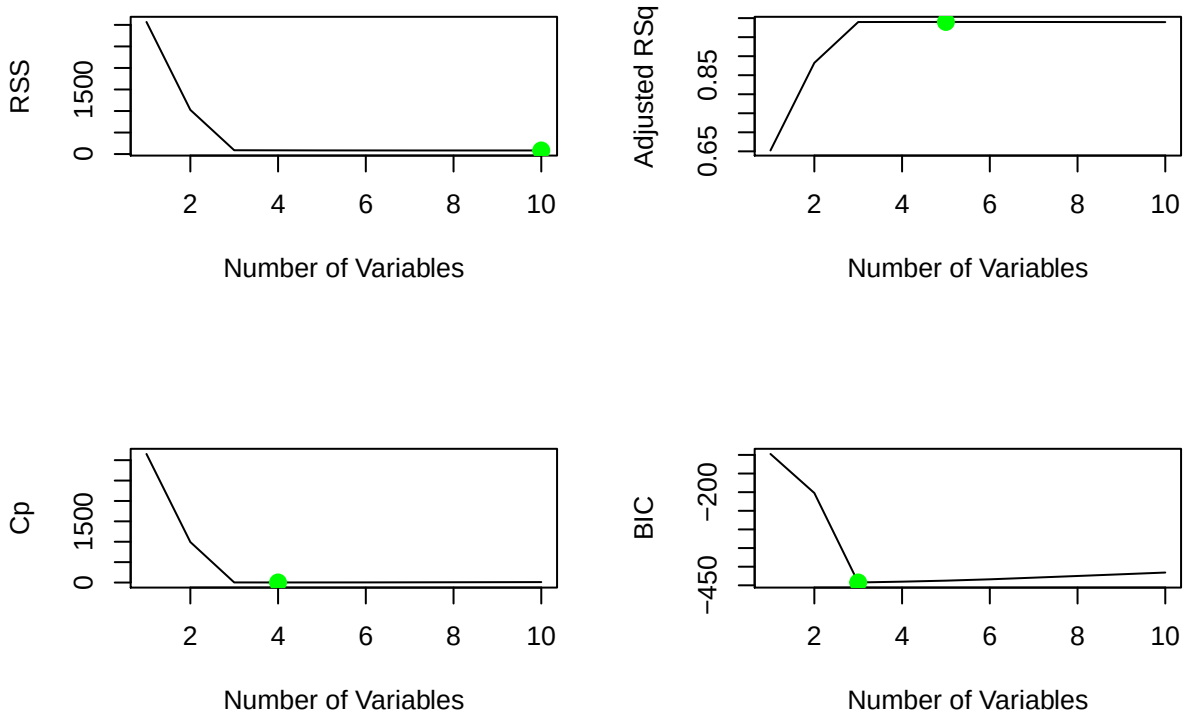
$$Y = 3.324088 + 76.459510x + 30.621314x^2 + 45.183672x^3$$

D

Forward Stepwise Selection:

```
foward<-regsubsets(Y~poly(X,10), data=frame, nvmax = 10,method="forward")
summary <- summary(foward)
par(mfrow = c(2,2))
plot(summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")
rss_min <- which.min(summary$rss)
points(rss_min, summary$rss[rss_min], col = "green", cex = 2, pch = 20)

plot(summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adjr2_max <- which.max(summary$adjr2)
points(adjr2_max, summary$adjr2[adjr2_max], col = "green", cex = 2, pch = 20)
plot(summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
cp_min <- which.min(summary$cp)
points(cp_min, summary$cp[cp_min], col = "green", cex = 2, pch = 20)
plot(summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min <- which.min(summary$bic)
points(bic_min, summary$bic[bic_min], col = "green", cex = 2, pch = 20)
```




```
coef(foward, which.max(summary$adjr2))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)4
##     3.324088    76.459510    30.621314    45.183672     1.257095
## poly(X, 10)5
##     1.480188
```

```
coef(foward, which.min(summary$bic))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3
```

```
##     3.324088    76.459510    30.621314    45.183672
```

```
coef(foward, which.min(summary$cp))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)5
##     3.324088    76.459510    30.621314    45.183672     1.480188
```

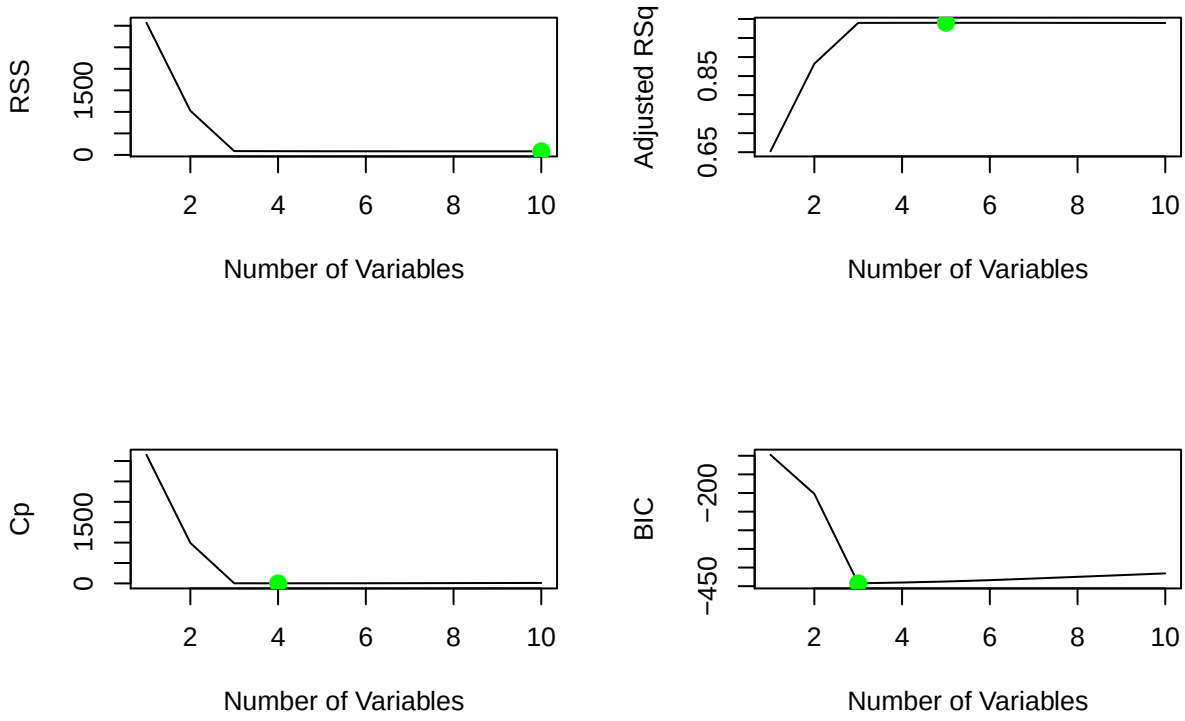
We see that **forward selection**, using Cp we will choose the 4-variables model. The best model obtained according to Cp is the model that uses X^1, X^2, X^3, and X^5 as predictors. And using BIC we pick the model that uses three variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3 as predictors. Adjusted R^2 tells us to pick model that has 5-variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3, X^4, and X^5 as predictors. RSS tells us to choose the full ten variable model.After looking at the scatterplot of our data it looks to be a cubic relationship(we also know that the underlying relationship is cubic plus an error) so the best model would be the model obtained by BIC:

$$Y = 3.324088 + 76.459510x + 30.621314x^2 + 45.183672x^3$$

Backwards Stepwise Selection:

```
backward<-regsubsets(Y~poly(X,10), data=frame, nvmax = 10,method="backward")
summary <- summary(backward)
par(mfrow = c(2,2))
plot(summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")
rss_min <- which.min(summary$rss)
points(rss_min, summary$rss[rss_min], col = "green", cex = 2, pch = 20)

plot(summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")
adjr2_max <- which.max(summary$adjr2)
points(adjr2_max, summary$adjr2[adjr2_max], col = "green", cex = 2, pch = 20)
plot(summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
cp_min <- which.min(summary$cp)
points(cp_min, summary$cp[cp_min], col = "green", cex = 2, pch = 20)
plot(summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
bic_min <- which.min(summary$bic)
points(bic_min, summary$bic[bic_min], col = "green", cex = 2, pch = 20)
```




```
coef(best, which.max(summary$adjr2))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)4
##     3.324088    76.459510    30.621314    45.183672     1.257095
## poly(X, 10)5
##     1.480188
```

```
coef(best, which.min(summary$bic))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3
##     3.324088    76.459510    30.621314    45.183672
```

```
coef(best, which.min(summary$cp))
```

```
##  (Intercept) poly(X, 10)1 poly(X, 10)2 poly(X, 10)3 poly(X, 10)5
##     3.324088    76.459510    30.621314    45.183672     1.480188
```

We see that **backwards selection**, using Cp we will choose the 4-variables model. The best model obtained according to Cp is the model that uses X^1, X^2, X^3, and X^5 as predictors. And using BIC we pick the model that uses three variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3 as predictors. Adjusted R^2 tells us to pick model that has 5-variables.The best model obtained according to BIC is the model that uses X^1, X^2, X^3, X^4, and X^5 as predictors. RSS tells us to choose the full ten variable model.After looking at the scatterplot of our data it looks to be a cubic relationship(we also know that the underlying relationship is cubic plus an error) so the best model would be the model obtained by BIC:

$$Y = 3.324088 + 76.459510x + 30.621314x^2 + 45.183672x^3$$

Our answers for both backwards selection and foward selection happen to match with the results found using best subset selection (in c).