Individual assignment5

Yuhan_Xu

2019/9/30

Problem 8

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a)

Q: Use the rnorm() function to generate a predictor X of length n = 100, as well as a noise vector of length n = 100.

A:

```
set.seed(1)
X = rnorm(100)
noise = rnorm(100)
```

(b)

Q: Generate a response vector Y of length n = 100 according to the model

$Y = \beta 0 + \beta 1X + \beta 2X2 + \beta 3X3 + \epsilon$

Where β0, β1, β2, and β3 are constants of your choice.

A:
When $\beta 0 = 4, \beta 1 = 3, \beta 2 = -2, \beta 3 = 1$:

```
Y = 4 + 3*X - 2*X^2 + 1*X^3 + noise
Y

##   [1]   0.4695349   4.5317895  -1.3978566   7.9138929   4.1525644
##   [6]   1.4072320   5.8196273   6.4373797   5.6393665   4.5510056
##  [11]   6.7837916   4.4631770   2.5569193 -23.9674589   6.0600387
##  [16]   3.4682625   3.6309078   5.6115363   6.1628785   5.1084161
##  [21]   5.3380301   6.9444352   3.9984102 -17.9355514   5.2290457
##  [26]   4.5378024   3.4067231  -7.9575194   1.3173162   4.6532081
##  [31]   6.9523151   3.0805257   5.4521954   2.3142451  -6.2285153
##  [36]   0.8026549   2.1439268   3.2865351   5.5589618   5.5122580
##  [41]   1.5334804   4.2718502   3.7929573   4.7592068  -0.4576910
##  [46]  -0.2285395   6.9635327   5.5956372   2.3499995   4.1340647
##  [51]   5.3906233   1.1669578   4.5122590  -4.3088320   5.6472927
##  [56]   8.7891361   3.5791407  -3.0724383   3.8604910   5.4251840
```

```
## [61]  13.9463917   3.6404929   6.5043572   4.9688828  -0.3645978
## [66]   6.7079233 -14.0659850   5.8242614   4.2719866  11.5401421
## [71]   7.3898055   0.6100852   5.7709973  -1.4395586  -5.2083006
## [76]   4.6944866   2.9776384   6.0785586   5.2397740   2.5393965
## [81]   0.2320037   4.5393431   6.6134618 -10.2170503   5.8068448
## [86]   4.6552941   7.5950198   2.1081650   4.4566776   3.7515586
## [91]   1.4470012   6.8699382   5.6189131   6.2937300   6.5120475
## [96]   4.1778566  -3.7284408   0.4186958  -4.0977375   1.6444128
```

(c)

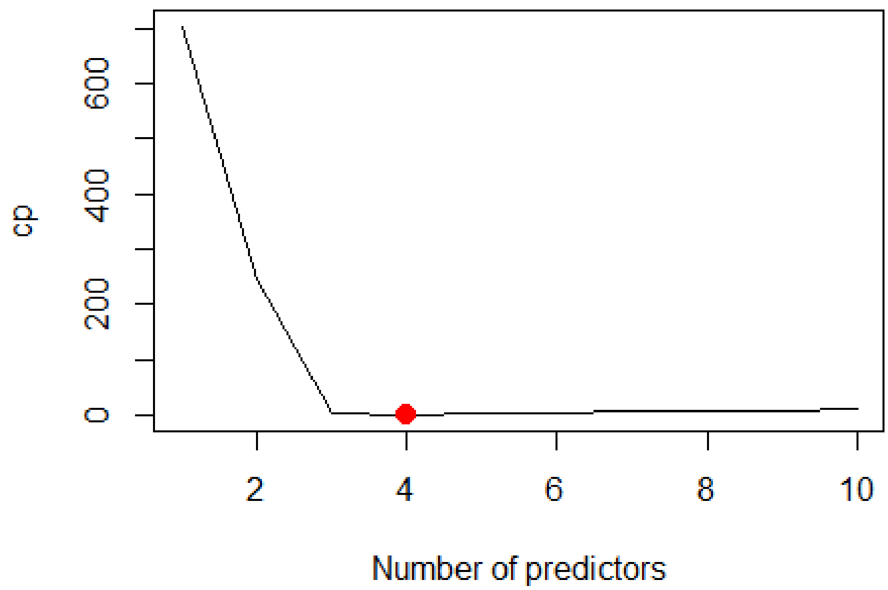
Q: Use the regsubsets() function to perform best subset selection in order to choose the best model containing the predictors X, X2, ..., X10. What is the best model obtained according to Cp, BIC, and adjusted R2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the data.frame() function to create a single data set containing both X and Y.

A:

```
xydata = data.frame(Y, X)

library("leaps")
regfit.full =regsubsets(Y~poly(X, 10, raw = TRUE), data = xydata, nvmax
 = 10)
reg.summary = summary(regfit.full)
names(reg.summary)

## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "
obj"
```

Cp

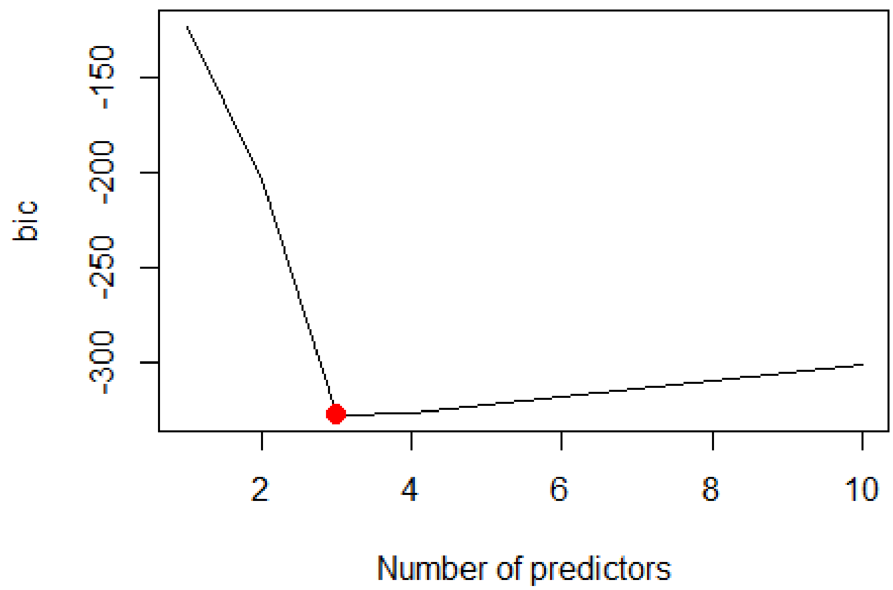
```
plot(reg.summary$cp, xlab = "Number of predictors", ylab = "cp", type =
 "l")
which.min(reg.summary$cp)

## [1] 4

points(4, reg.summary$cp[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the smallest Cp.

BIC

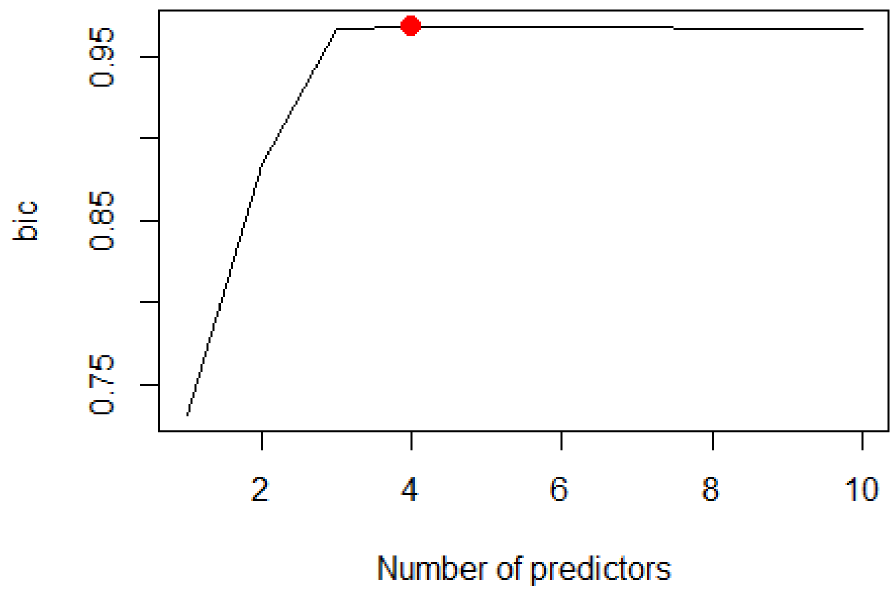
```
plot(reg.summary$bic, xlab = "Number of predictors", ylab = "bic", type
 = "l")
which.min(reg.summary$bic)

## [1] 3

points(3, reg.summary$bic[3], col = "red", cex = 2, pch = 20)
```



When there are 3 predictors, we can get the model with the smallest BIC.

Adjusted R2

```
plot(reg.summary$adjr2, xlab = "Number of predictors", ylab = "bic", ty
pe = "l")
which.max(reg.summary$adjr2)

## [1] 4

points(4, reg.summary$adjr2[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the largest adjusted R2.

Even though comparing adjusted R2 value as well as BIC we will get the best model with 4 predictors, from the plot we can see that after the number of predictors increases to 3, the value of Cp, BIC and adjusted R2 all changed dramatically. Therefore, the best model obtained according to these 3 values is the 3-predictor model. The coefficients are:

```
coef(regfit.full, 3)

##                (Intercept) poly(X, 10, raw = TRUE)1 poly(X, 10, raw = 
TRUE)2
##                   4.061507                 2.975280                -2.
123791
## poly(X, 10, raw = TRUE)3
##                   1.017639
```

This model also agrees with the original true function $Y = 4 + 3X - 2X^2 + X^3 + \epsilon$.

(d)

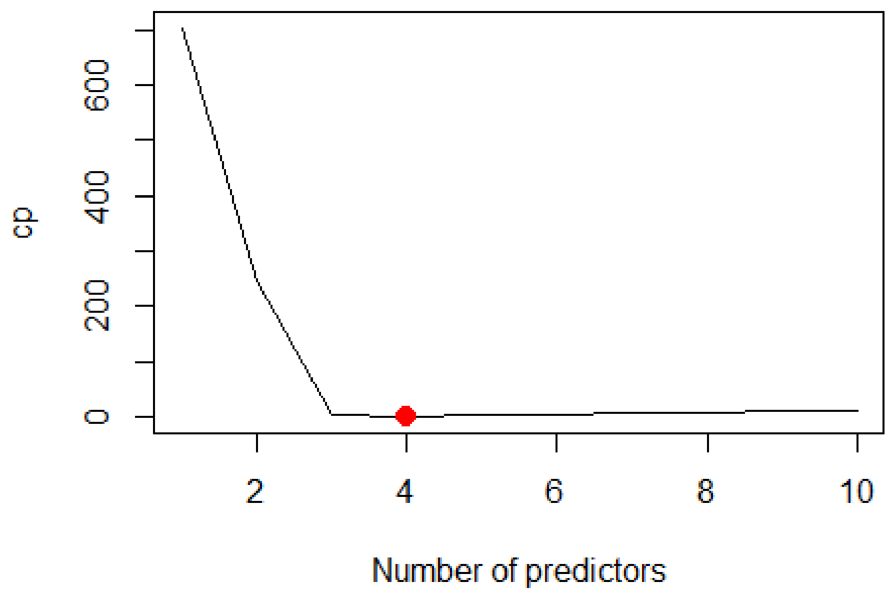
Q: Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

A: Using forward stepwise selection:

```r
regfit.fwd = regsubsets(Y~poly(X, 10, raw = TRUE), data = xydata, nvmax
 = 10, method = "forward")
reg.summary1 = summary(regfit.fwd)
```

Cp

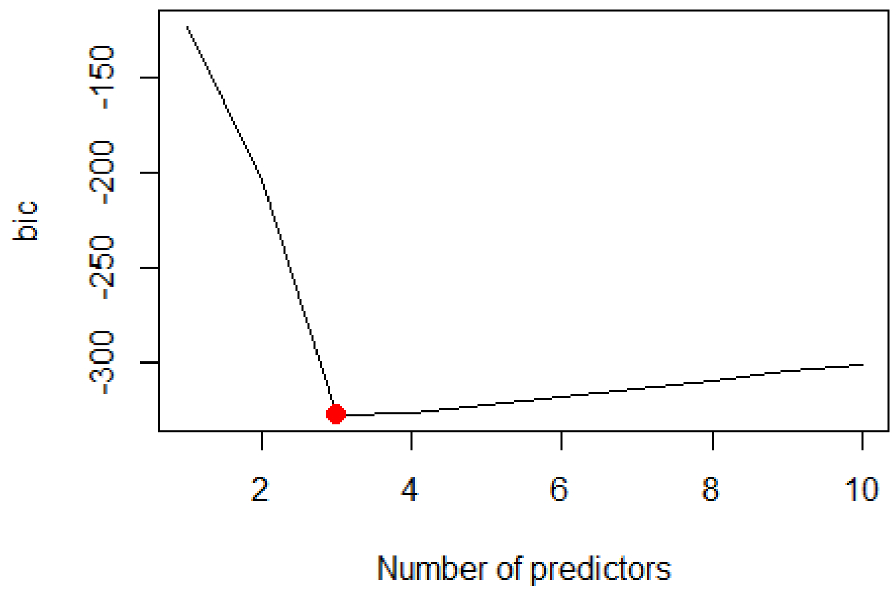
```r
plot(reg.summary1$cp, xlab = "Number of predictors", ylab = "cp", type
= "l")
which.min(reg.summary1$cp)

## [1] 4

points(4, reg.summary1$cp[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the smallest Cp.

BIC

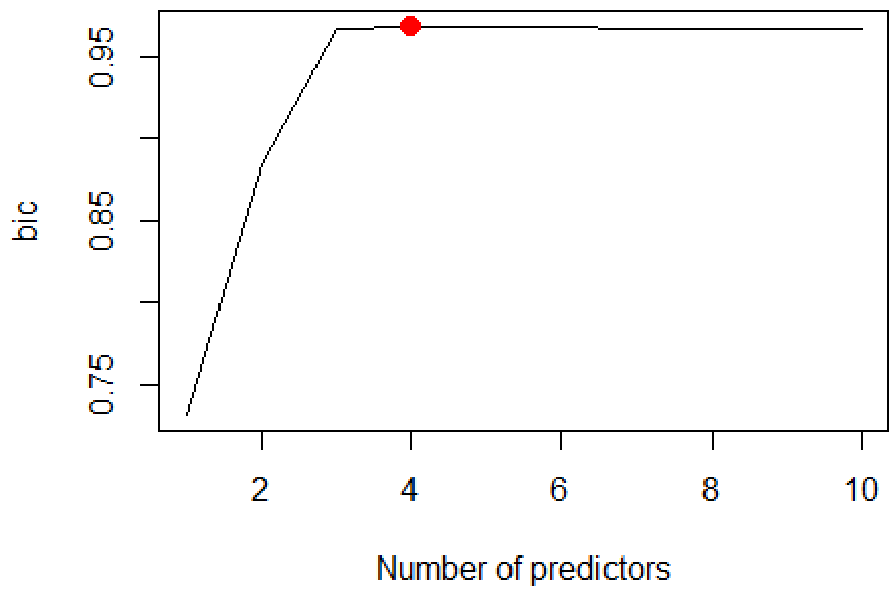
```r
plot(reg.summary1$bic, xlab = "Number of predictors", ylab = "bic", typ
e = "l")
which.min(reg.summary1$bic)

## [1] 3

points(3, reg.summary1$bic[3], col = "red", cex = 2, pch = 20)
```



When there are 3 predictors, we can get the model with the smallest BIC.

Adjusted R2

```
plot(reg.summary1$adjr2, xlab = "Number of predictors", ylab = "bic", t
ype = "l")
which.max(reg.summary1$adjr2)

## [1] 4

points(4, reg.summary1$adjr2[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the largest adjusted R2.

Similarly, the model obtained according to the value of Cp, BIC, adjusted R2 is also the 3-predictor model. The coefficients are:

```
coef(regfit.fwd, 3)

##                (Intercept) poly(X, 10, raw = TRUE)1 poly(X, 10, raw = 
TRUE)2
##                   4.061507                 2.975280                -2.
123791
## poly(X, 10, raw = TRUE)3 
##                   1.017639
```

These coefficients are the same as those in the best model obtained using best subset method.

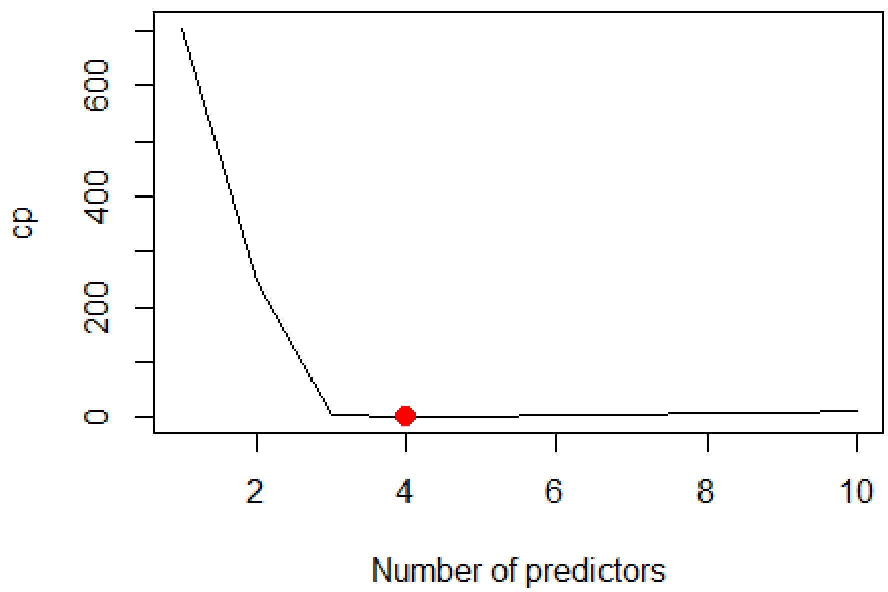
Using backward stepwise selection:

```
regfit.bwd = regsubsets(Y~poly(X, 10, raw = TRUE), data = xydata, nvmax
 = 10, method = "backward")
reg.summary2 = summary(regfit.bwd)
```

Cp

```
plot(reg.summary2$cp, xlab = "Number of predictors", ylab = "cp", type
= "l")
which.min(reg.summary2$cp)
```

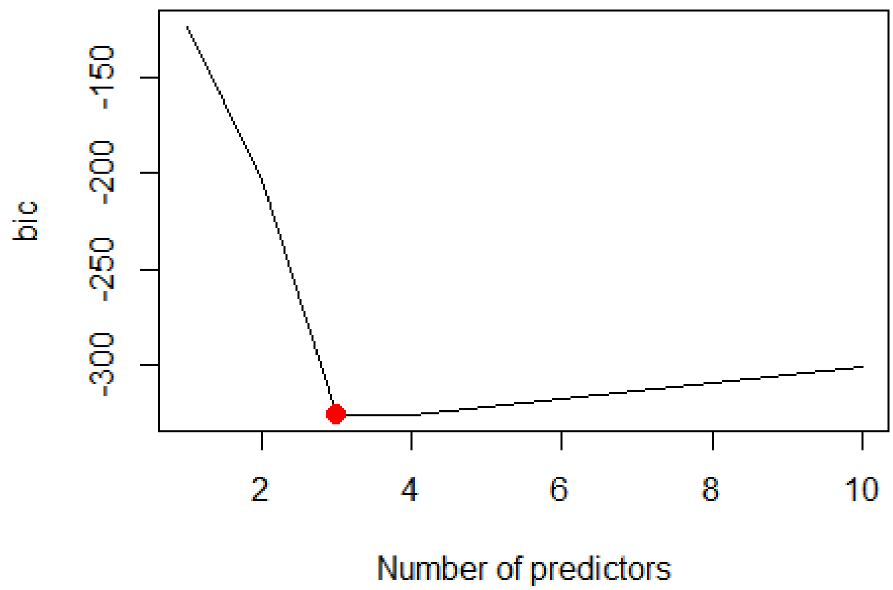
```
## [1] 4
points(4, reg.summary2$cp[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the smallest Cp.

BIC

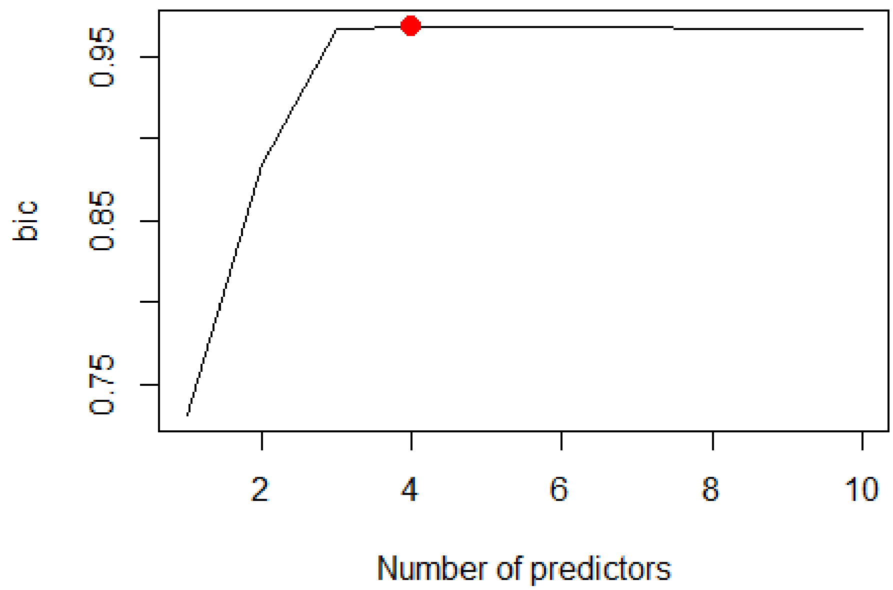
```
plot(reg.summary2$bic, xlab = "Number of predictors", ylab = "bic", typ
e = "l")
which.min(reg.summary2$bic)

## [1] 3

points(3, reg.summary2$bic[3], col = "red", cex = 2, pch = 20)
```



When there are 3 predictors, we can get the model with the smallest Cp.

Adjusted R2

```
plot(reg.summary2$adjr2, xlab = "Number of predictors", ylab = "bic", t
ype = "l")
which.max(reg.summary2$adjr2)

## [1] 4

points(4, reg.summary2$adjr2[4], col = "red", cex = 2, pch = 20)
```



When there are 4 predictors, we can get the model with the largest adjusted R2.

Similarly, the best model should be the 3-predictor model. The coefficients of the best model are:

```
coef(regfit.bwd, 3)

##              (Intercept) poly(X, 10, raw = TRUE)1 poly(X, 10, raw = 
TRUE)2
##                4.0738073                3.9427063               -2.1
784855
## poly(X, 10, raw = TRUE)5
##                0.1721833
```

The best model obtained using backward stepwise method is a little bit different from the model obtained by other two methods. Backward stepwise method chooses X^5 rather than X^3, and the coefficients of 3 predictors and intercept are also different.