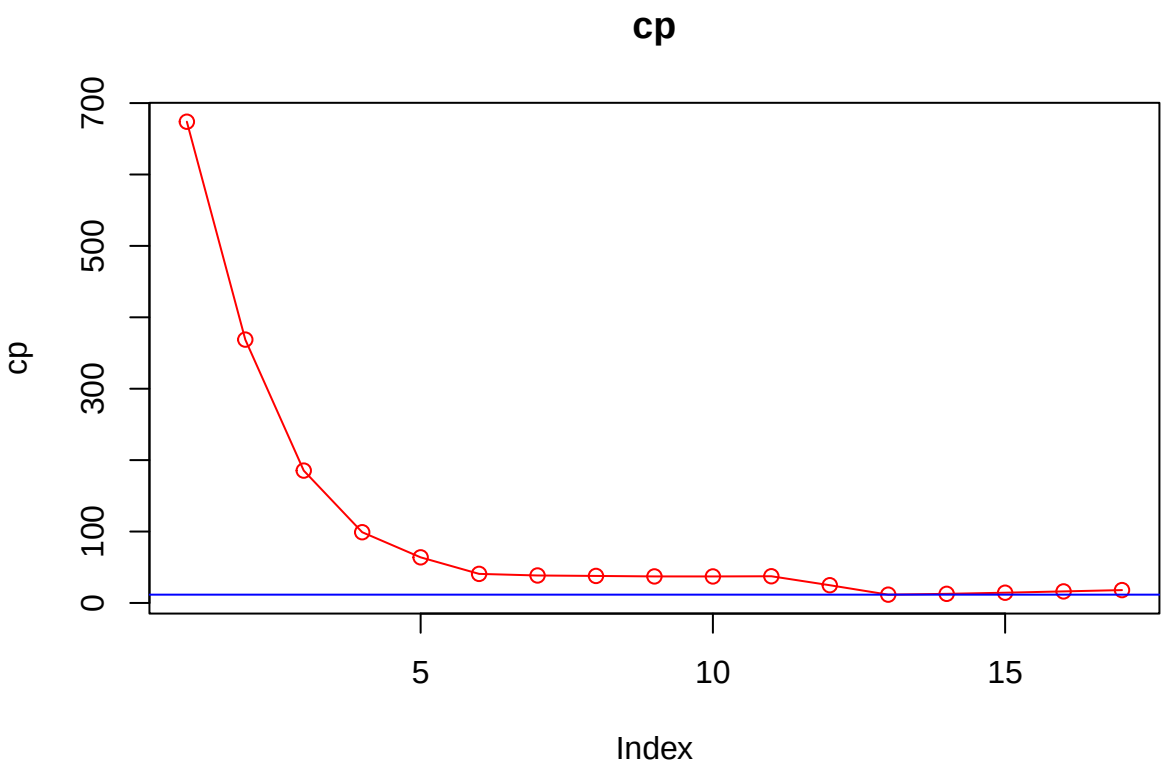
A4Q2

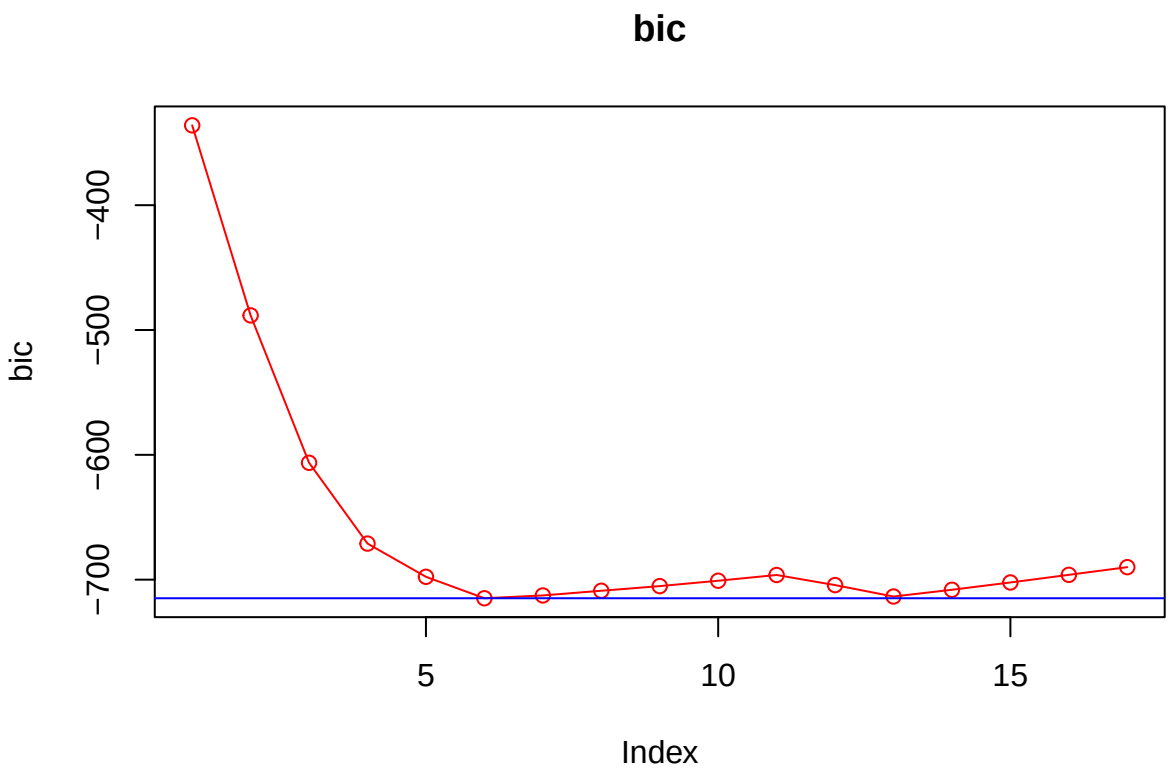
Undergraduate Student

(i)

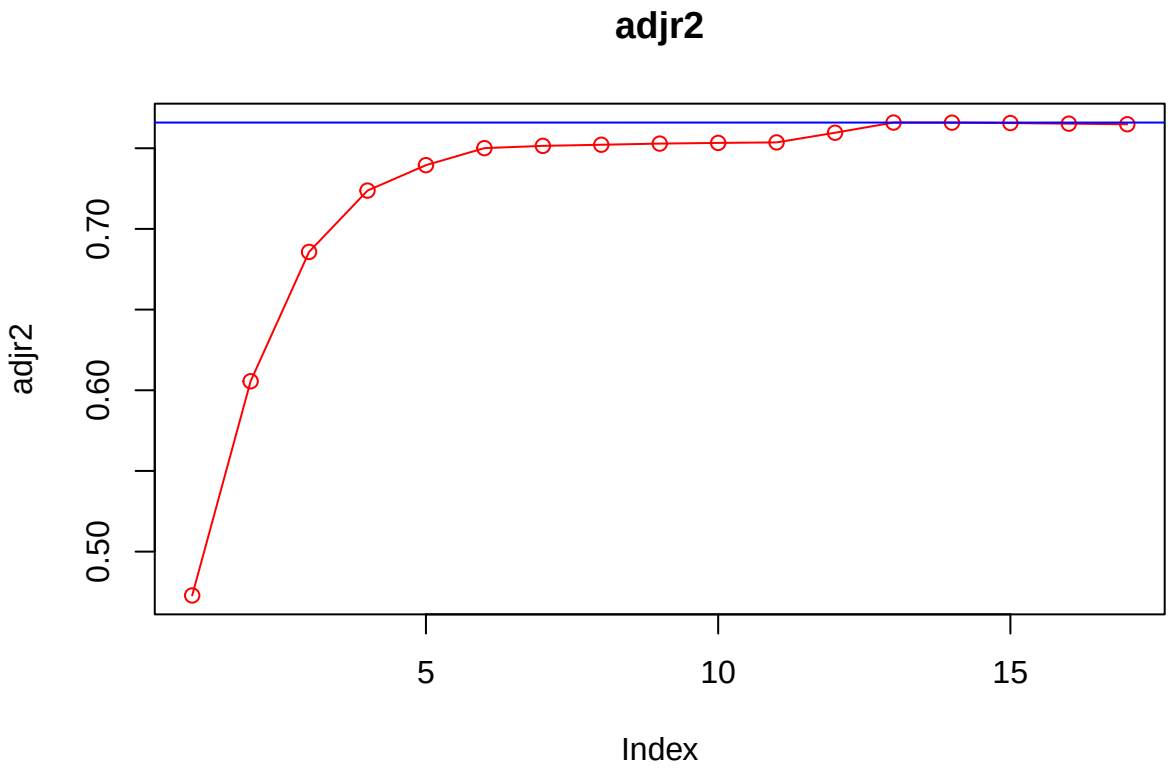
```
library(leaps)
set.seed(6)
college <- read.csv("College.csv", header=T)[,-c(1)]
college$Private <- as.numeric(as.factor(college$Private))
s <- sample(nrow(college), nrow(college)*0.7)
train <- college[s,]
test <- college[-s,]
step <-regsubsets(Outstate~., data=train, method="forward", nvmax=ncol(train)-1)
subsummary <- summary(step)
plot(subsummary$cp,type="o",col="red",main="cp",ylab="cp")
abline(h=min(subsummary$cp),col="blue")
```



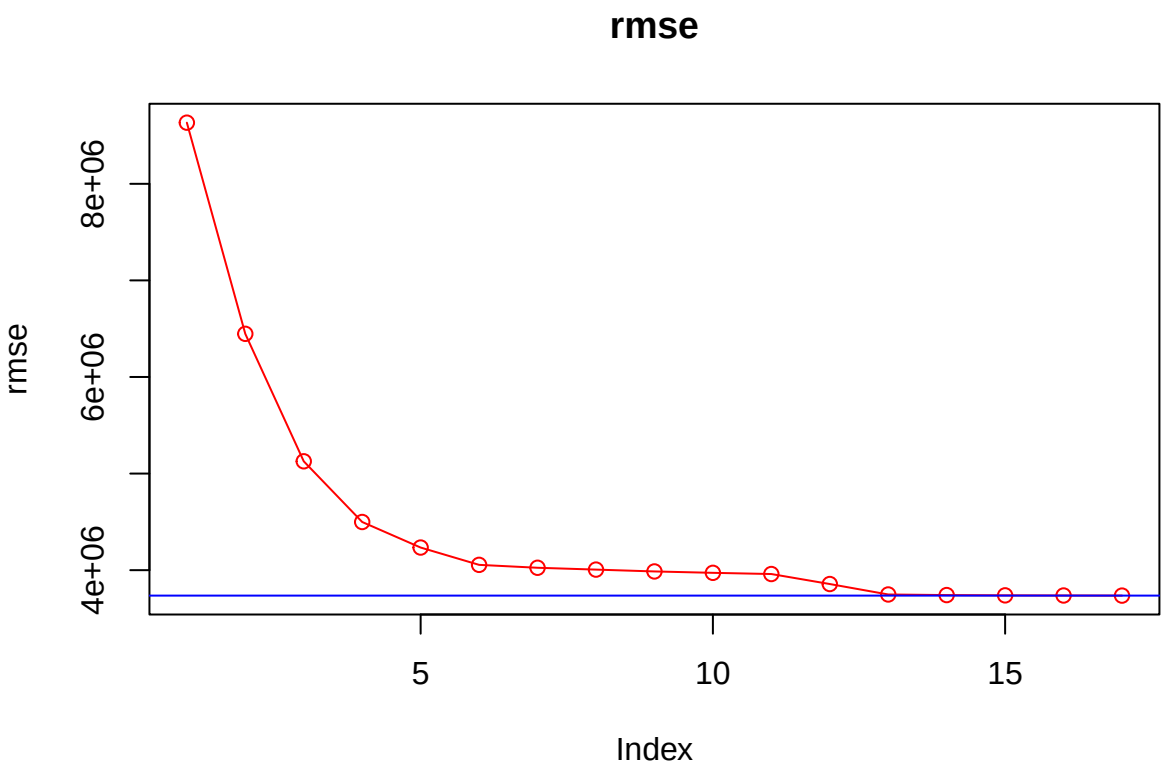

```
plot(subsummary$bic,type="o",col="red",main="bic",ylab="bic")
abline(h=min(subsummary$bic),col="blue")
```




```
plot(subsummary$adjr2,type="o",col="red",main="adjr2", ylab="adjr2")
abline(h=max(subsummary$adjr2),col="blue")
```



```
plot((subsummary$rss/nrow(train)),type="o",col="red",main="rmse", ,ylab="rmse")
abline(h=min(subsummary$rss/nrow(train)),col="blue")
```




```
which.min(subsummary$cp)
```

```
## [1] 13
```

```
which.min(subsummary$bic)
```

```
## [1] 6
```

```
which.max(subsummary$adjr2)
```

```
## [1] 13
```

```
which.min(subsummary$rss/nrow(train))
```

```
## [1] 17
```

6 is the best minimum size we can get from four graphs.

```
subsummary
```

```
## Subset selection object
## Call: regsubsets.formula(Outstate ~ ., data = train, method = "forward",
##     nvmax = ncol(train) - 1)
## 17 Variables  (and intercept)
##             Forced in Forced out
## Private         FALSE      FALSE
## Apps            FALSE      FALSE
## Accept          FALSE      FALSE
## Enroll          FALSE      FALSE
## Top10perc       FALSE      FALSE
## Top25perc       FALSE      FALSE
## F.Undergrad     FALSE      FALSE
## P.Undergrad     FALSE      FALSE
## Room.Board      FALSE      FALSE
## Books           FALSE      FALSE
## Personal        FALSE      FALSE
## PhD             FALSE      FALSE
## Terminal        FALSE      FALSE
## S.F.Ratio       FALSE      FALSE
## perc.alumni     FALSE      FALSE
## Expend          FALSE      FALSE
## Grad.Rate       FALSE      FALSE
## 1 subsets of each size up to 17
## Selection Algorithm: forward
##           Private Apps Accept Enroll Top10perc Top25perc F.Undergrad
## 1  ( 1 )  " "     " "  " "    " "    " "        " "        " "
## 2  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 3  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 4  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 5  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 6  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 7  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 8  ( 1 )  "*"     " "  " "    " "    " "        " "        " "
## 9  ( 1 )  "*"     " "  " "    " "    "*"        " "        " "
## 10  ( 1 ) "*"     " "  " "    " "    "*"        " "        " "
## 11  ( 1 ) "*"     " "  "*"    " "    "*"        " "        " "
## 12  ( 1 ) "*"     "*"  "*"    " "    "*"        " "        " "
## 13  ( 1 ) "*"     "*"  "*"    " "    "*"        " "        "*"
## 14  ( 1 ) "*"     "*"  "*"    " "    "*"        "*"        "*"
## 15  ( 1 ) "*"     "*"  "*"    "*"    "*"        "*"        "*"
## 16  ( 1 ) "*"     "*"  "*"    "*"    "*"        "*"        "*"
## 17  ( 1 ) "*"     "*"  "*"    "*"    "*"        "*"        "*"
##           P.Undergrad Room.Board Books Personal PhD Terminal S.F.Ratio
## 1  ( 1 )  " "         " "        " "   " "      " " " "      " "
## 2  ( 1 )  " "         " "        " "   " "      " " " "      " "
## 3  ( 1 )  " "         "*"        " "   " "      " " " "      " "
## 4  ( 1 )  " "         "*"        " "   " "      " " " "      " "
## 5  ( 1 )  " "         "*"        " "   " "      " " "*"      " "
## 6  ( 1 )  " "         "*"        " "   " "      " " "*"      " "
## 7  ( 1 )  " "         "*"        " "   " "      " " "*"      "*"
## 8  ( 1 )  " "         "*"        " "   "*"      " " "*"      "*"
## 9  ( 1 )  " "         "*"        " "   "*"      " " "*"      "*"
## 10  ( 1 ) " "         "*"        "*"   "*"      " " "*"      "*"
## 11  ( 1 ) " "         "*"        "*"   "*"      " " "*"      "*"
```

```
## 12  ( 1 ) " "         "*"         "*"    "*"       " " "*"       "*"
## 13  ( 1 ) " "         "*"         "*"    "*"       " " "*"       "*"
## 14  ( 1 ) " "         "*"         "*"    "*"       " " "*"       "*"
## 15  ( 1 ) " "         "*"         "*"    "*"       " " "*"       "*"
## 16  ( 1 ) "*"         "*"         "*"    "*"       " " "*"       "*"
## 17  ( 1 ) "*"         "*"         "*"    "*"       "*" "*"       "*"
##           perc.alumni Expend Grad.Rate
## 1  ( 1 )  " "         "*"     " "
## 2  ( 1 )  " "         "*"     " "
## 3  ( 1 )  " "         "*"     " "
## 4  ( 1 )  "*"         "*"     " "
## 5  ( 1 )  "*"         "*"     " "
## 6  ( 1 )  "*"         "*"     "*"
## 7  ( 1 )  "*"         "*"     "*"
## 8  ( 1 )  "*"         "*"     "*"
## 9  ( 1 )  "*"         "*"     "*"
## 10  ( 1 ) "*"         "*"     "*"
## 11  ( 1 ) "*"         "*"     "*"
## 12  ( 1 ) "*"         "*"     "*"
## 13  ( 1 ) "*"         "*"     "*"
## 14  ( 1 ) "*"         "*"     "*"
## 15  ( 1 ) "*"         "*"     "*"
## 16  ( 1 ) "*"         "*"     "*"
## 17  ( 1 ) "*"         "*"     "*"
```

We get Private, Room.Board, Expend, Grad.Rate, perc.alumni, Terminal

(ii)

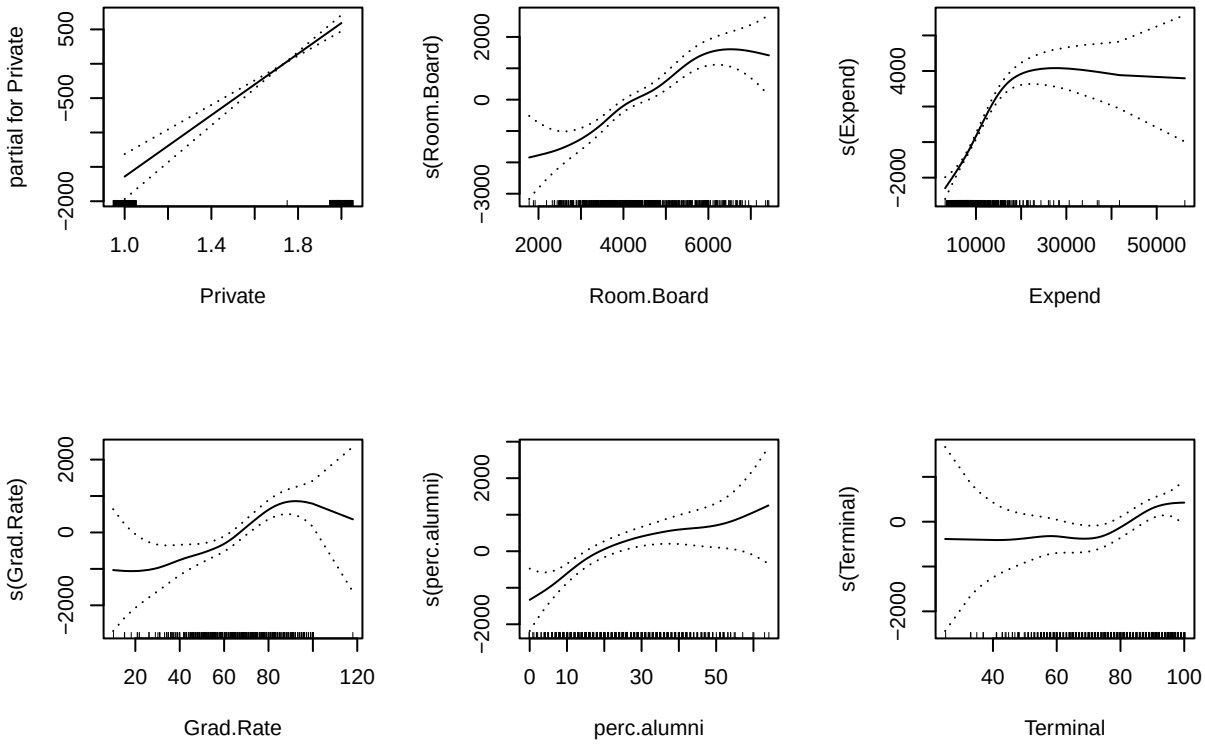
```
library(gam)
```

```
## Loading required package: splines
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.20.1
```

```
gamfit <- gam(Outstate~Private+s(Room.Board)+s(Expend)+s(Grad.Rate)+
              s(perc.alumni)+s(Terminal),data = train)
par(mfrow=c(2,3))
plot(gamfit, se=T)
```



Comments: From graphs, we see that there might exist non-linear relationship between OutState and Room.Board, OutState and Expend.

(iii)

```
pred <- predict(gamfit, test)
mse <- mean((test$Outstate-pred)^2)
tss <- mean((test$Outstate-mean(test$Outstate))^2)
R <- 1 - mse/tss
mse
```

```
## [1] 3627254
```

```
R
```

```
## [1] 0.7667036
```

Comments: After applying the model to test data set, we get mean square error 3627254 and R squared 0.7667036. This shows that this model does not perform extremely well.

(iv)

```
summary(gamfit)
```

```
##
## Call: gam(formula = Outstate ~ Private + s(Room.Board) + s(Expend) +
##     s(Grad.Rate) + s(perc.alumni) + s(Terminal), data = train)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -6864.87 -1142.35    44.97  1248.98  7841.07
##
## (Dispersion Parameter for gaussian family taken to be 3419787)
##
##     Null Deviance: 8908512519 on 542 degrees of freedom
## Residual Deviance: 1781709096 on 521 degrees of freedom
## AIC: 9733.993
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                  Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Private           1 2166885318 2166885318 633.6316 < 2.2e-16 ***
## s(Room.Board)     1 2094996326 2094996326 612.6101 < 2.2e-16 ***
## s(Expend)         1 1323903791 1323903791 387.1304 < 2.2e-16 ***
## s(Grad.Rate)      1  233644547  233644547  68.3214 1.160e-15 ***
## s(perc.alumni)    1  100901382  100901382  29.5052 8.571e-08 ***
## s(Terminal)       1   27554367   27554367   8.0573  0.004709 **
## Residuals       521 1781709096    3419787
## ---
## 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)         3  3.3955 0.01775 *
## s(Expend)             3 30.1071 < 2e-16 ***
## s(Grad.Rate)          3  2.2165 0.08529 .
## s(perc.alumni)        3  1.9444 0.12141
## s(Terminal)           3  1.6524 0.17637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: From Anova Test for Nonparametric Effects, we see that the p-value of Expend is extremely small and Room.Board is smaller than 0.05. This means that there is a really strong evidence that Expend has a non-linear relationship with OutState, and there is an evidence that Room.Board might have non-linear relationship with OutState. This matches our conclusion in part ii.