Statistical Methods for Bioinformatics [I0U31a] Assignment 05 - Chapter 7

Hamed Borhani

May 17, 2016

7.9.10

(a)

```
library(ISLR)
library(leaps)
attach(College)

#train/testr
set.seed(2)
train <- sample(1:nrow(College), nrow(College)/2)

fwd.college <- regsubsets(Outstate~., data = College[train,],
                          nvmax = 17, method = "forward")
> plot(summary(fwd.college)$cp, type = 'b')
> plot(summary(fwd.college)$bic, type = 'b')
> plot(summary(fwd.college)$adjr2, type = 'b')
```

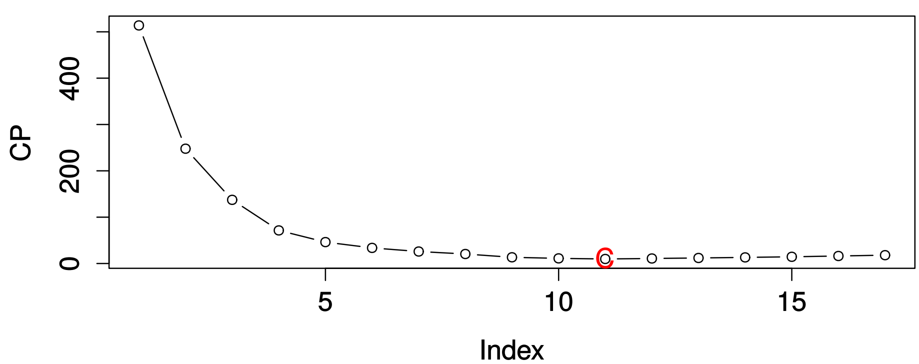


Figure 1: CP plot

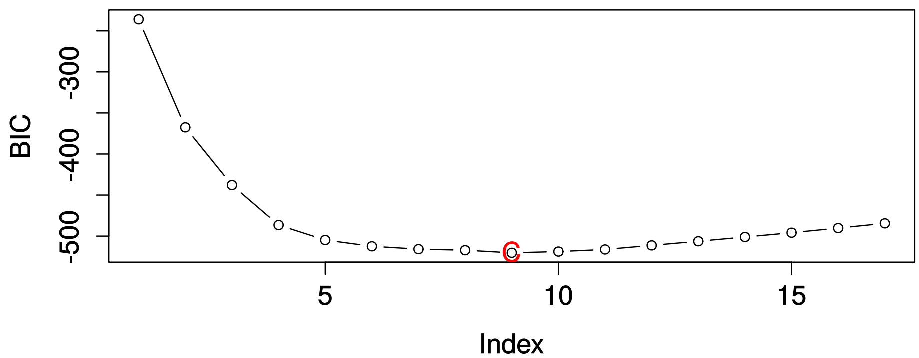


Figure 2: BIC plot

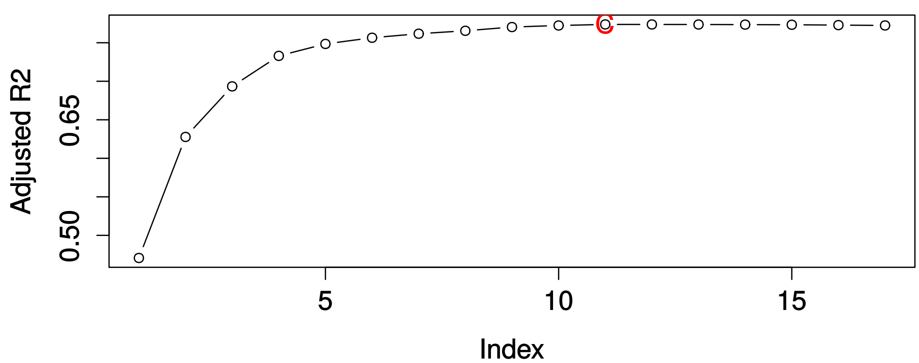


Figure 3: Adjusted R^2

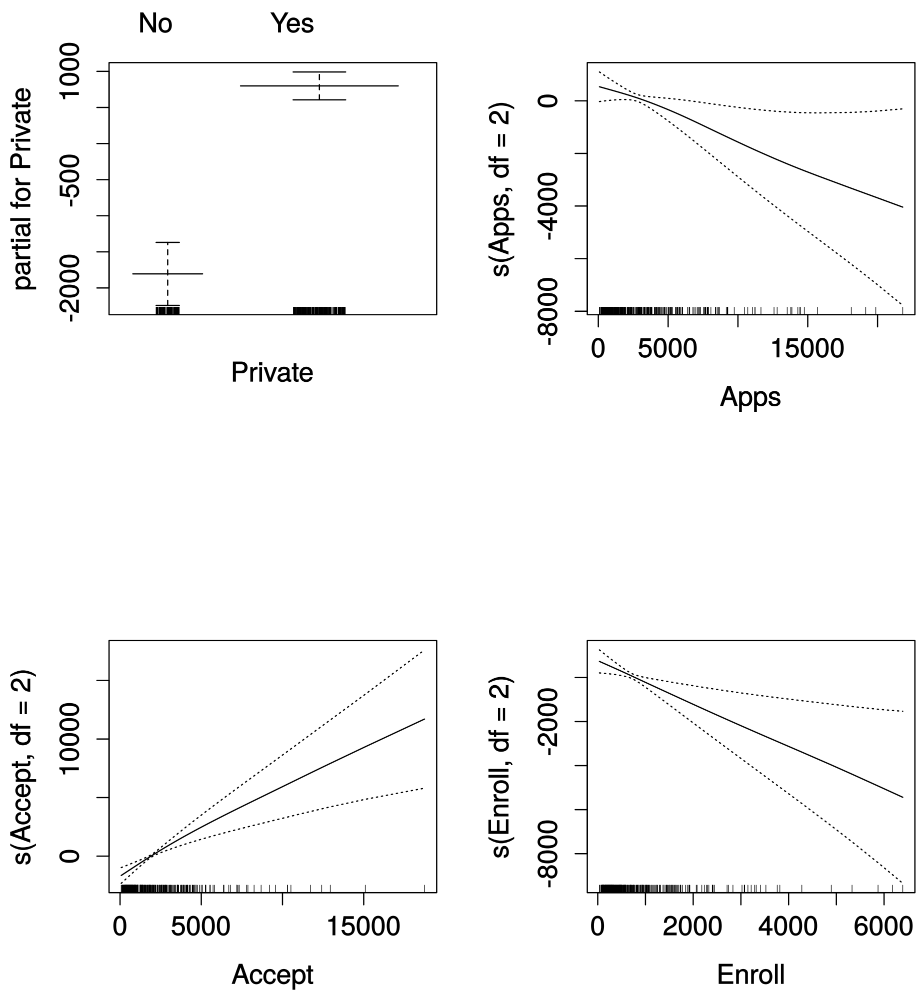
Minimum CP and also maximum adjusted R^2 correspond to the model with 11 variables. But minimum BIC corresponds to the model with 9 variables. So we fit a model with this 11 variables:

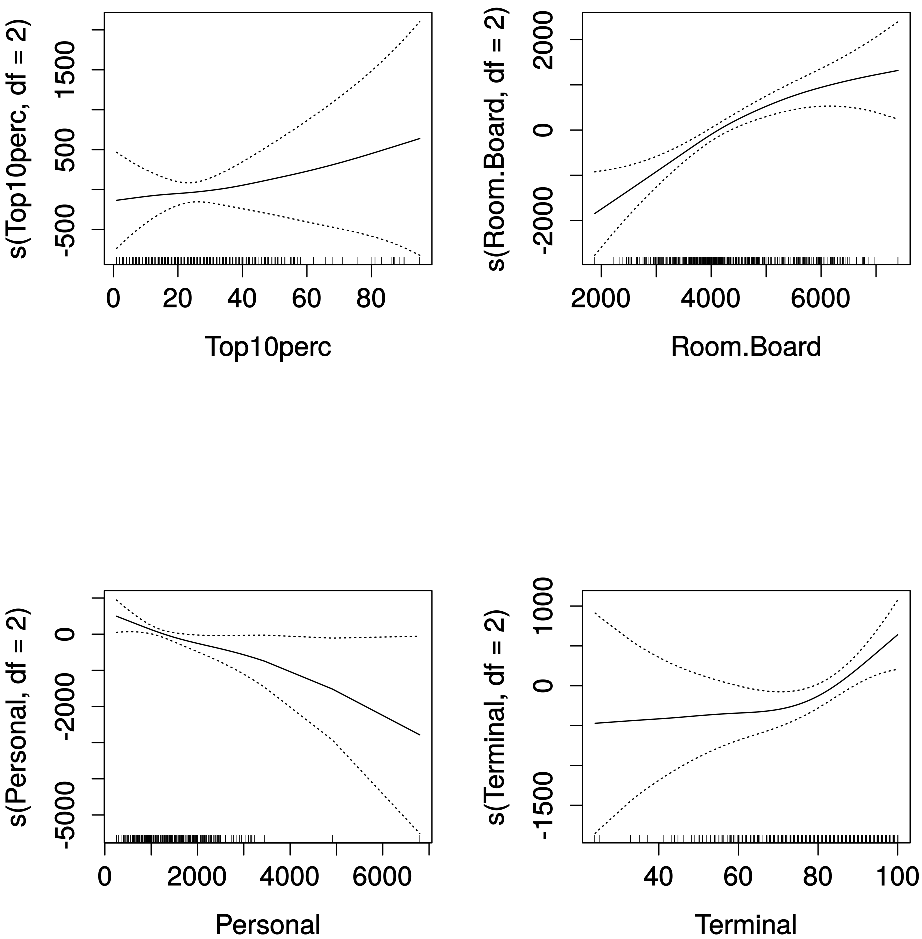
```
> coef(fwd.college, id = 11)
```

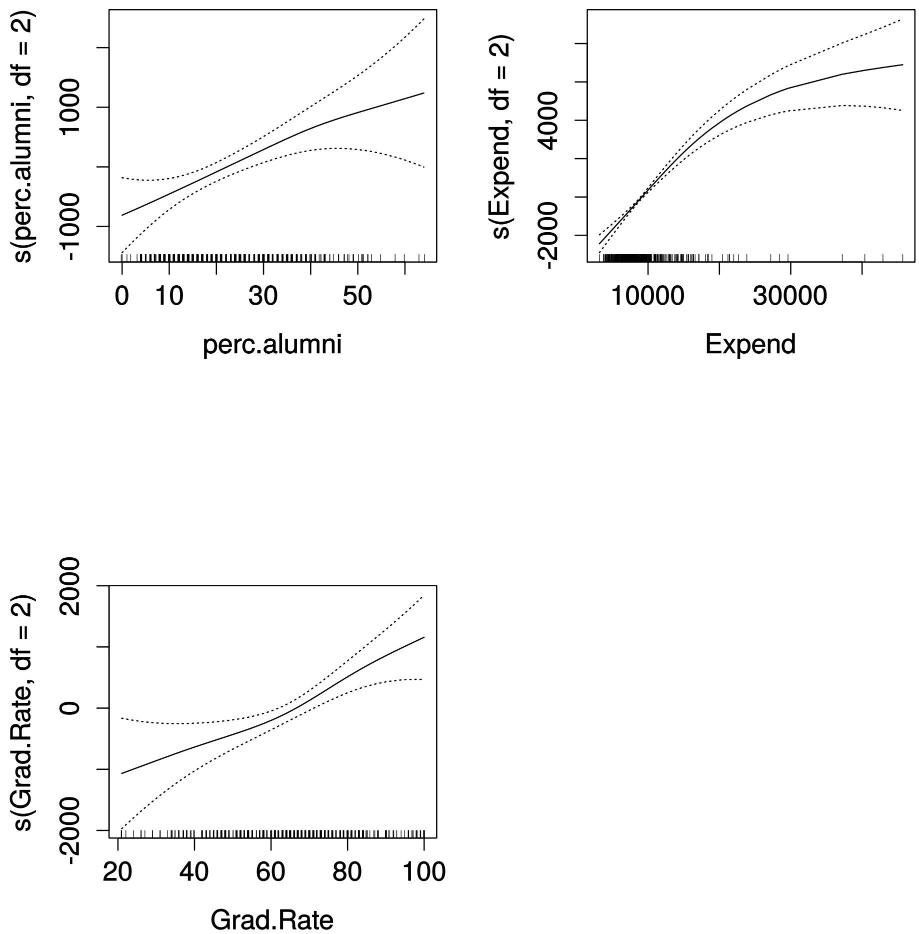
```
  (Intercept)    PrivateYes          Apps         Accept
      Enroll
-2266.1474174  2591.6010582    -0.2576897      0.8452096
    -1.0789110
     Top10perc    Room.Board      Personal       Terminal     perc.
       alumni
   17.3710739     0.7653948    -0.3793180     27.0520396
      37.8893471
       Expend     Grad.Rate
    0.2642780    28.2822643

lm.college <- lm(Outstate~ Private + Apps + Accept + Enroll
                 + Top10perc + Room.Board + Personal + Terminal
                 + perc.alumni + Expend + Grad.Rate,
                 data= College, subset = train)

yhat.lm <- predict(lm.college, newdata = College[-train,])

> min((yhat.lm - College[-train, "Outstate"])^2)
[1] 0.1008338
```

(b)

```
library(gam)

gam.college <- gam(Outstate~ Private + s(Apps, df=2) + s(Accept,
     df=2)
                   + s(Enroll, df=2) + s(Top10perc, df=2)
                   + s(Room.Board, df=2) + s(Personal, df=2)
                   + s(Terminal, df=2) + s(perc.alumni, df=2)
                   + s(Expend, df=2) + s(Grad.Rate, df=2), data=
                        College, subset = train)

> plot(gam.college, se=TRUE)
```


No
Yes
partial for Private
1000
-500
-2000
Private
s(Apps, df = 2)
0
-4000
-8000
0
5000
15000
Apps
s(Accept, df = 2)
10000
0
0
5000
15000
Accept
s(Enroll, df = 2)
-2000
-8000
0
2000
4000
6000
Enroll


s(Top10perc, df = 2)
Top10perc
s(Room.Board, df = 2)
Room.Board
s(Personal, df = 2)
Personal
s(Terminal, df = 2)
Terminal



(c)

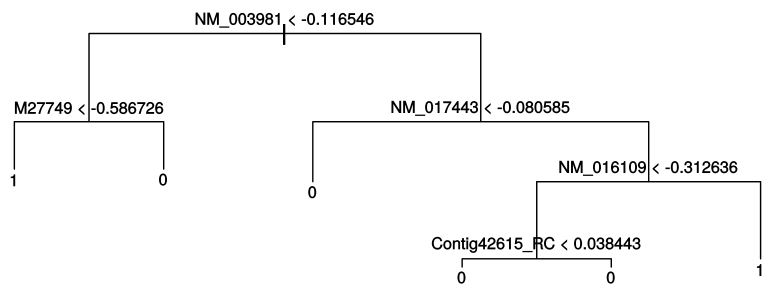
```
yhat.gam <- predict(gam.college, newdata = College[-train,])
> min((yhat.gam - College[-train, "Outstate"])^2)
```

```
[1] 1.739645e-05
```

Test error for the GAM fit is very small, which shows that it's a good fit.

Vervij Data

```
library(tree)

#dummy coding : DM->1 , NODM->0
phenotype = ifelse(data$meta == "DM", 1, 0)
vijver <- cbind(phenotype,data[,-1])

vijver$phenotype <- as.factor(vijver$phenotype)

#train/test
train <- sample(1:nrow(vijver), nrow(vijver)/2)
vijver.test <- vijver[-train, "phenotype"]

#building a tree
vijver.tree <- tree(phenotype~., data = vijver[train,])

> summary(vijver.tree)
Classification tree:
tree(formula = phenotype ~ ., data = vijver[train, ])
Variables actually used in tree construction:
[1] "NM_003981"        "M27749"           "NM_017443"        "NM_
    016109"
[5] "Contig42615_RC"
Number of terminal nodes:  6
Residual mean deviance:  0.1633 = 14.37 / 88
Misclassification error rate: 0.04255 = 4 / 94

tree.pred <- predict(vijver.tree, newdata = vijver[-train,],
    type = 'class')
> table(tree.pred, vijver.test)
         vijver.test
tree.pred  0  1
        0 40 23
        1 14 17
```


NM_003981 < -0.116546
M27749 < -0.586726
NM_017443 < -0.080585
NM_016109 < -0.312636
Contig42615_RC < 0.038443
1
0
0
0
0
1