BestSubsetSelection

Kiranmayi

March 25, 2019

Explaining Best Subset Selection (BSS) through an example:

The dataset I considered is Credit dataset from ISLR package. The dataset contains information on ten thousand customers to predict which customers will default on their credit card debt. The continuous response is Balance which indicates credit card balance is dollars. The 6 predictors, I am considering to fit the model are described below:
Income - A quantitative variable indicating income in $10000s
Limit - credit limit, quantitative variable
Rating - credit rating, quantitative variable
Cards - Number of credit cards
Age - Age in years
Education - Number of years in education

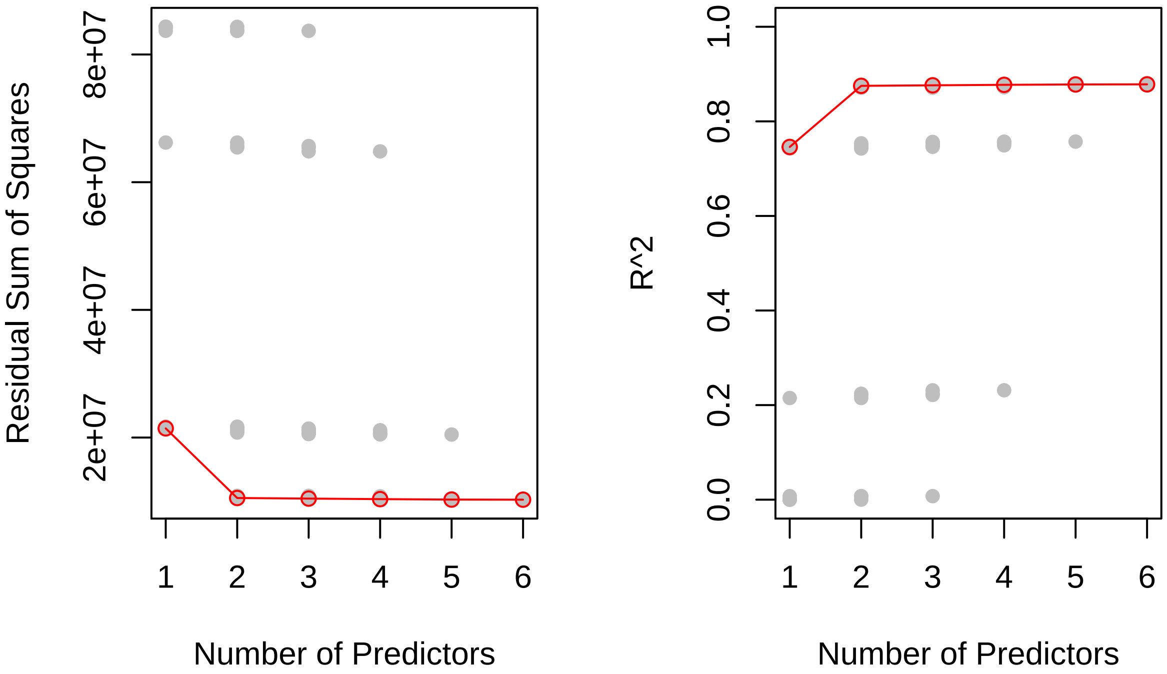
Fitting a least squares linear model using every possible subset of features,

```r
library(ISLR)
library(plyr)
library(leaps)

credit = data.frame(balance = Credit$Balance,income = Credit$Income,limit=Credit$Limit,
                    rating=Credit$Rating,cards=Credit$Cards,age=Credit$Age,
                    education=Credit$Education)
n = dim(credit)[1]
features = names(credit)[2:7]
#generating all possible subsets
subsets = c(1)
for (i in 1:length(features)){
  x= combn(features,i,simplify = F)
  subsets = c(subsets,x)
}
ex = paste('balance~',sapply(subsets,paste,collapse='+'))
#subsets
#length(subsets)

#Fitting least squares model to every subset and storing the results in outputs dataframe
coeff = list()
outputs = data.frame(rss=numeric(),sig=numeric(),rsq=numeric(),adjr=numeric(),fealen=numeric())
for (j in 1:length(subsets)){
  mod = lm(as.formula(ex[j]), data=credit)
  sm = summary(mod)
  outputs[j,'rss'] = sum(sm$residuals^2)
  outputs[j,'fealen']= length(subsets[[j]])
  outputs[j,'sigma']=sm$sigma
  outputs[j,'rsq']=sm$r.squared
  outputs[j,'adjr']=sm$adj.r.squared
  coeff[[j]] = sm$coefficients
}
```

All possible subsets of 6 features are generated and fitted least squares linear model using each subset. Number of models for BSS = $2^6 = 64$

Plotting RSS vs number of predictors and R^2 vs number of predictors. The plots will contain these values corresponding to each subset and then the best models for number of predictors 1,..6 are displayed

```
#calculating minimum RSS and maximum R^2
min.rss = tapply(outputs$rss,outputs$fealen,min)
max.rsq = tapply(outputs$rsq,outputs$fealen,max)
#plotting the graphs
par(mfrow=c(1,2))
plot(outputs$fealen,outputs$rss,col='grey',pch=16,xlab='Number of Predictors'
     ,ylab='Residual Sum of Squares')
points(min.rss,type='o',col='red')
plot(outputs$fealen,outputs$rsq,col='grey',pch=16, ylim=c(0,1),
     xlab='Number of Predictors',ylab='R^2')
points(max.rsq,type='o',col='red')
```



Displaying the predictors in each of the models M_0, M_1,..M_6

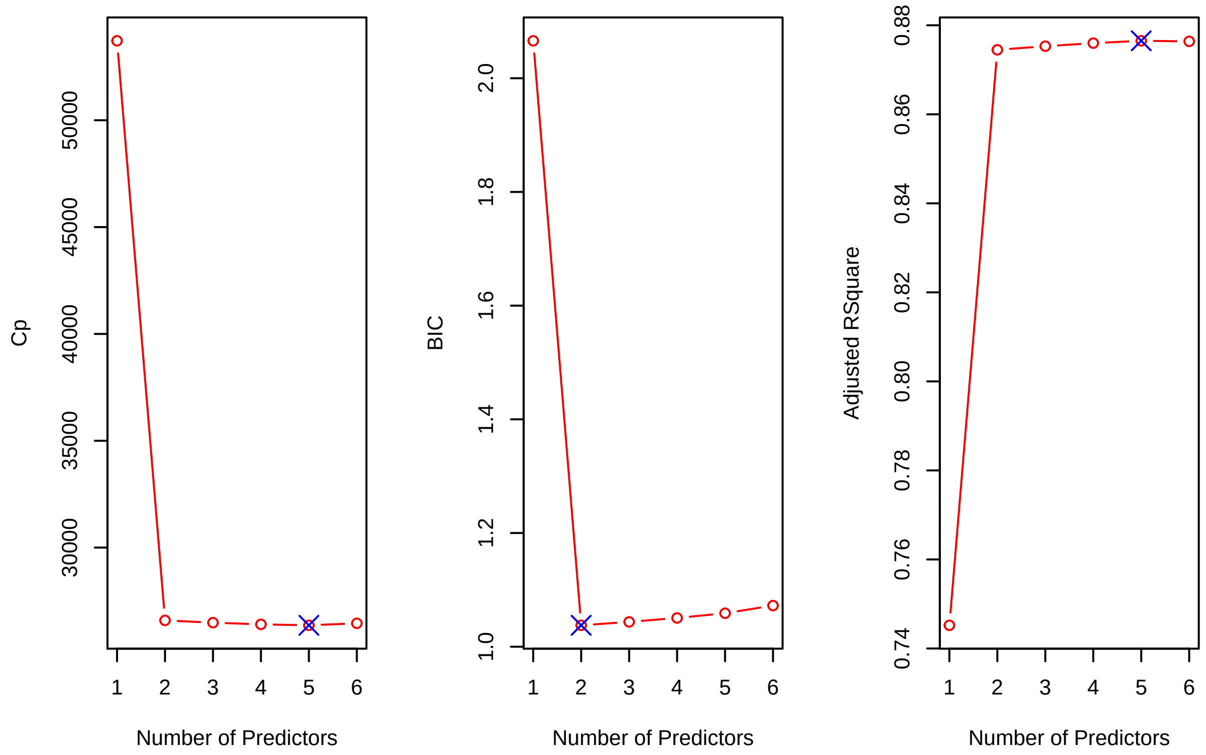
```
outputs$subsets=subsets
M = data.frame(adjrs = outputs[which(outputs$rss %in% min.rss),'adjr'],
               fealens=outputs[which(outputs$rss %in% min.rss),'fealen'],
               rss = min.rss)
M$preds = outputs[which(outputs$rss %in% min.rss),'subsets']
M$preds
```

```
## [[1]]
## [1] "rating"
##
## [[2]]
## [1] "income" "rating"
##
## [[3]]
## [1] "income" "limit"  "rating"
##
## [[4]]
## [1] "income" "limit"  "rating" "age"
##
## [[5]]
## [1] "income" "limit"  "rating" "cards"  "age"
##
## [[6]]
## [1] "income"    "limit"     "rating"    "cards"     "age"       "education"
```

The predictors are:
M_0 - This is a null model. It does not contain any predictors.
M_1 - The best model for one predictor uses “rating” as predictor.
M_2 - The best model for two predictors uses “income” and “rating” as predictors.
M_3 - The best model for three predictors uses “income”, “limit” and “rating” as predictors.
M_4 - The best model for four predictors uses “income”, “limit”, “rating” and “age” as predictors.
M_5 - The best model for five predictors uses “income”, “limit”, “rating”, “cards” and “age” as predictors.
M_6 - This is the full model and contains all predictors “income”, “limit”, “rating”, “cards”, “age” and “education”.

Calculating Cp, BIC and adjusted R^2 for the models M_0,. . . M_6 and plotting them vs number of predictors. The optimal model in each case are also displayed.

```r
sigsq = outputs[64,'sigma']^2
#function to calculate Cp
cp = function(r,d){
  (r+2*d*sigsq)/n
}
M$cp = mapply(cp,M$rss,M$fealens)
#function to calculate BIC
bic = function(r,d){
  (r+log(n)*d*sigsq)/(n*sigsq)
}
M$bic = mapply(bic,M$rss,M$fealens)
#plotting the graphs
par(mfrow=c(1,3))
plot(M$fealens,M$cp, col='red',type='b', xlab='Number of Predictors',ylab='Cp')
bestc = which.min(M$cp)
points (bestc, M$cp[bestc], col ="blue",cex =2, pch =4)
plot(M$fealens,M$bic, col='red',type='b', xlab='Number of Predictors',ylab='BIC')
bestb = which.min(M$bic)
points (bestb, M$bic[bestb], col ="blue",cex =2, pch =4)
plot(M$fealens,M$adjrs, col='red',type='b', xlab='Number of Predictors',ylab='Adjusted RSquare')
besta = which.max(M$adjrs)
points (besta, M$adjrs[besta], col ="blue",cex =2, pch =4)
```



From the graphs above, considering Cp and Adjusted R^2 indicate that model with five predictors is the best one and BIC indicates that model with 2 predictors is the best one.

So following the majority, I am considering the 5 predictors model as the best one. Calculating coefficient estimates for this model.

```
bestrss.c= M[bestc,'rss']
posc = which(outputs$rss==bestrss.c)
subsets[posc]
```

```
## [[1]]
## [1] "income" "limit"  "rating" "cards"  "age"
```

```
coeff[posc]
```

```
## [[1]]
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) -449.3610111 40.57408781 -11.075074 5.303629e-25
## income        -7.5621098  0.38213548 -19.789080 5.143084e-61
## limit          0.1285533  0.05289267   2.430456 1.552541e-02
## rating         2.0224047  0.79207704   2.553293 1.104684e-02
## cards         11.5527211  7.06284686   1.635703 1.027003e-01
## age           -0.8883161  0.47780635  -1.859155 6.375045e-02
```

```r
bestadjr = max(M$adjrs)
posa = which(outputs$adjr==bestadjr)
subsets[posa]
```

```
## [[1]]
## [1] "income" "limit"  "rating" "cards"  "age"
```

```r
coeff[posa]
```

```
## [[1]]
##                  Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) -449.3610111 40.57408781 -11.075074 5.303629e-25
## income        -7.5621098  0.38213548 -19.789080 5.143084e-61
## limit          0.1285533  0.05289267   2.430456 1.552541e-02
## rating         2.0224047  0.79207704   2.553293 1.104684e-02
## cards         11.5527211  7.06284686   1.635703 1.027003e-01
## age           -0.8883161  0.47780635  -1.859155 6.375045e-02
```

The best model through Cp and Adjusted R^2 are same as displayed above. This model contains five predictors "income", "limit", "rating", "cards" and "age".
The intercept of the model is -449.361. This can be understood as that when all predictors are zero, estimated balance is $-449.361.
The coefficient of income is -7.56. With $10000 increase in income, the average balance decreases by $7.56 considering all other variables to remain same.
The coefficient of limit is 0.129. With one unit increase in credit limit, the average balance increases by $0.129 considering all other variables to remain same.
The coefficient of rating is 2.022. With one unit increase in rating, the average average balance increases by $2.022 considering all other variables to remain same.
The coefficient of cards is 11.553. With one number increase in cards, the average balance increases by $11.553 considering all other variables to remain same.
The coefficient of age is -0.888. With one year increase in age, the average balance decreases by $0.888 considering all other variables to remain same.

Checking the best model through BIC,

```r
bestrss.b = M[bestb,'rss']
posb = which(outputs$rss==bestrss.b)
subsets[posb]
```

```
## [[1]]
## [1] "income" "rating"
```

```r
coeff[posb]
```

```
## [[1]]
##                 Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) -534.812150 21.60269845 -24.75673  1.663589e-82
## income        -7.672124  0.37846203 -20.27185  3.107102e-63
## rating         3.949265  0.08620904  45.81034 1.448189e-160
```

If we consider best model through BIC, the model contains only two predictors “income” and “rating”.
The intercept of the model is -534.821. This can be understood as that when all predictors are zero, estimated balance is $-534.821.
The coefficient of income is -7.67. With $10000 increase in income, the average balance decreases by $7.67 considering all other variables to remain same.
The coefficient of rating is 3.949. With one unit increase in rating, the average balance increases by $3.949 considering all other variables to remain same.