NBA 4920/6921 Lecture 10

Linear Model Best Subset Selection Application

Murat Unal

9/30/2021

```r
rm(list=ls())
options(digits = 3, scipen = 999)
library(tidyverse)
library(ISLR)
library(cowplot)
library(ggcorrplot)
library(stargazer)
library(corrr)
library(lmtest)
library(sandwich)
library(MASS)
library(car)
library(jtools)
library(caret)
library(leaps)
library(future.apply)
hitters <- ISLR::Hitters
hitters <- na.omit(hitters)
set.seed(2)
```

```
dim(hitters)
```

```
[1] 263  20
```

```
names(hitters)
```

```
 [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"
 [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns
[13] "CWalks"    "League"    "Division"  "PutOuts"   "Assis
[19] "Salary"    "NewLeague"
```

Best subset selection

```r
# Draw validation set
hit_validation_data = hitters %>% sample_frac(size = 0.3)
# Create the remaining training set
hit_training_data = setdiff(hitters, hit_validation_data)
```

```r
nvars = 19
regfit.best=regsubsets(Salary~.,data=hit_training_data,
                                              nvmax=nvars)
best.sum <- summary(regfit.best)
best.model <- which.max(best.sum$adjr2)
best.model
```

```
[1] 10
```

```r
coef(regfit.best,id=best.model)[1:4]
```

```
(Intercept)        AtBat         Hits        Walks
      88.31        -1.69         6.05         5.59
```

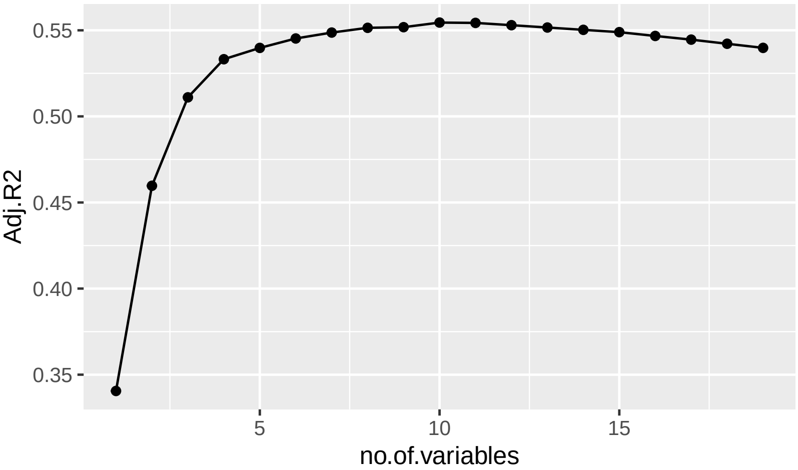
```r
coef(regfit.best,id=best.model)[5:9]
```

```
CAtBat CHmRun  CRuns   CRBI CWalks
-0.130 -1.868  1.448  1.204 -0.912
```

```r
coef(regfit.best,id=best.model)[10:11]
```

```
DivisionW    PutOuts
  -87.035      0.233
```

```r
plot.adjr2 <- data.frame("no.of.variables"=seq(1:nvars),
                         "Adj.R2"= best.sum$adjr2)
ggplot(plot.adjr2,aes(x=no.of.variables,y=Adj.R2))+
  geom_point()+
  geom_line()
```

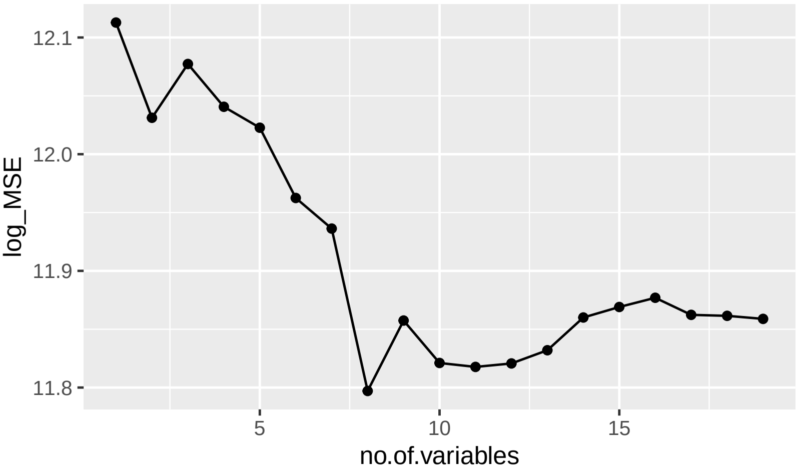

0.55
0.50
0.45
0.40
0.35
Adj.R2
5
10
15
no.of.variables

Validation set approach

```
validation.mat=model.matrix(Salary~.,
                            data=hit_validation_data)

val.errors = numeric(nvars)
for(each in 1:nvars){
    coefi = coef(regfit.best,id=each)
    pred = validation.mat[,names(coefi)]%*%coefi
    val.errors[each]=
      mean((hit_validation_data$Salary-pred)^2)
}

which.min(val.errors)
```

```
[1] 8
```

```r
plot.data <- data.frame("no.of.variables"=seq(1:nvars),
                        "log_MSE"=log(val.errors))

ggplot(plot.data,aes(x=no.of.variables,y=log_MSE))+
  geom_point()+
  geom_line()
```


12.1
12.0
11.9
11.8
log_MSE
5
10
15
no.of.variables

K-fold cross validation

```
nvars = 19
nfold = 10
# Create folds
fold.list <- createFolds(rownames(hitters),nfold)
# Empty vector to store the resulting MSEs
cv.errors =matrix(0,nfold,nvars,
                dimnames =list(NULL,paste (1:nvars)))

for(each in 1:nfold){
 train <- hitters[-fold.list[[each]],]
 validate <- hitters[fold.list[[each]],]

 best.fit=regsubsets(Salary~.,data=train,nvmax =19)
 validation.mat=model.matrix(Salary~.,data=validate)

}
```

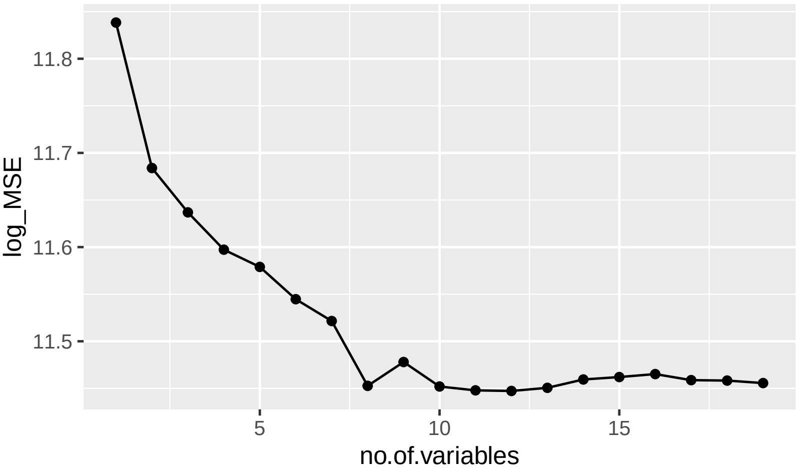
..continued from before

```
 for(i in 1:nvars){
   coefi = coef(regfit.best,id=i)
   pred = validation.mat[,names(coefi)]%*%coefi
   cv.errors[each,i] = mean( (validate$Salary-pred)^2)
   }
}
```

```
mean.cv.errors=apply(cv.errors ,2, mean)
which.min(mean.cv.errors)
```

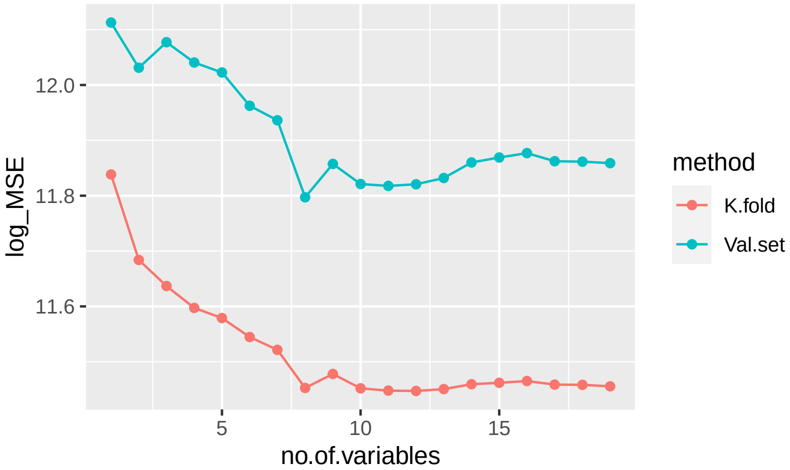
```
12
12
```

```
plot.data.fold <-data.frame(
                    "no.of.variables"=seq(1:nvars),
                    "log_MSE"=log(mean.cv.errors))

ggplot(plot.data.fold,aes(x=no.of.variables,y=log_MSE))+
  geom_point()+
  geom_line()
```



11.8
11.7
11.6
11.5
log_MSE
5
10
15
no.of.variables

```r
plot.data <- rbind(plot.data,plot.data.fold)
plot.data$method <- c(rep("Val.set",nvars),
                      rep("K.fold",nvars))

ggplot(plot.data,aes(x=no.of.variables,y=log_MSE,
                     color=method))+
  geom_point()+
  geom_line()
```


12.0
11.8
11.6
log_MSE
5
10
15
no.of.variables
method
K.fold
Val.set

To obtain the final model we perform best subset selection on the full data set and obtain the 8-variable model.

```
best.fit=regsubsets(Salary~.,data=hitters,nvmax =19)
coef(best.fit,8)[1:4]
```

```
(Intercept)       AtBat        Hits      Walks
     130.97       -2.17        7.36       6.00
```

```
coef(best.fit,8)[5:9]
```

```
   CHmRun      CRuns     CWalks DivisionW    PutOuts
    1.234      0.965     -0.832  -117.966      0.291
```

This is your final model that you'd deploy to predict the salary of baseball players.