Hitters-Subset-Selection-withAuto.R

duter

2021-02-17

#######################################################  
############# Example: Hitters -- Subset selection  
#######################################################  
  
## TASK 1 ##  
  
# load library ISLR and see what the dataframe Hitters looks like  
library(ISLR)  
View(Hitters)  
  
# check the names of the columns in Hitters  
names(Hitters)

## [1] "AtBat" "Hits" "HmRun" "Runs" "RBI" "Walks"   
## [7] "Years" "CAtBat" "CHits" "CHmRun" "CRuns" "CRBI"   
## [13] "CWalks" "League" "Division" "PutOuts" "Assists" "Errors"   
## [19] "Salary" "NewLeague"

# check the first five rows of Hitters  
Hitters[1:5, ]

## AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun  
## -Andy Allanson 293 66 1 30 29 14 1 293 66 1  
## -Alan Ashby 315 81 7 24 38 39 14 3449 835 69  
## -Alvin Davis 479 130 18 66 72 76 3 1624 457 63  
## -Andre Dawson 496 141 20 65 78 37 11 5628 1575 225  
## -Andres Galarraga 321 87 10 39 42 30 2 396 101 12  
## CRuns CRBI CWalks League Division PutOuts Assists Errors  
## -Andy Allanson 30 29 14 A E 446 33 20  
## -Alan Ashby 321 414 375 N W 632 43 10  
## -Alvin Davis 224 266 263 A W 880 82 14  
## -Andre Dawson 828 838 354 N E 200 11 3  
## -Andres Galarraga 48 46 33 N E 805 40 4  
## Salary NewLeague  
## -Andy Allanson NA A  
## -Alan Ashby 475.0 N  
## -Alvin Davis 480.0 A  
## -Andre Dawson 500.0 N  
## -Andres Galarraga 91.5 N

# check the dimensions of Hitters (how many rows, how many columns)  
dim(Hitters)

## [1] 322 20

# check the names of rows in Hitters  
rownames(Hitters)

## [1] "-Andy Allanson" "-Alan Ashby" "-Alvin Davis"   
## [4] "-Andre Dawson" "-Andres Galarraga" "-Alfredo Griffin"   
## [7] "-Al Newman" "-Argenis Salazar" "-Andres Thomas"   
## [10] "-Andre Thornton" "-Alan Trammell" "-Alex Trevino"   
## [13] "-Andy VanSlyke" "-Alan Wiggins" "-Bill Almon"   
## [16] "-Billy Beane" "-Buddy Bell" "-Buddy Biancalana"   
## [19] "-Bruce Bochte" "-Bruce Bochy" "-Barry Bonds"   
## [22] "-Bobby Bonilla" "-Bob Boone" "-Bob Brenly"   
## [25] "-Bill Buckner" "-Brett Butler" "-Bob Dernier"   
## [28] "-Bo Diaz" "-Bill Doran" "-Brian Downing"   
## [31] "-Bobby Grich" "-Billy Hatcher" "-Bob Horner"   
## [34] "-Brook Jacoby" "-Bob Kearney" "-Bill Madlock"   
## [37] "-Bobby Meacham" "-Bob Melvin" "-Ben Oglivie"   
## [40] "-Bip Roberts" "-BillyJo Robidoux" "-Bill Russell"   
## [43] "-Billy Sample" "-Bill Schroeder" "-Butch Wynegar"   
## [46] "-Chris Bando" "-Chris Brown" "-Carmen Castillo"   
## [49] "-Cecil Cooper" "-Chili Davis" "-Carlton Fisk"   
## [52] "-Curt Ford" "-Cliff Johnson" "-Carney Lansford"   
## [55] "-Chet Lemon" "-Candy Maldonado" "-Carmelo Martinez"   
## [58] "-Charlie Moore" "-Craig Reynolds" "-Cal Ripken"   
## [61] "-Cory Snyder" "-Chris Speier" "-Curt Wilkerson"   
## [64] "-Dave Anderson" "-Doug Baker" "-Don Baylor"   
## [67] "-Dann Bilardello" "-Daryl Boston" "-Darnell Coles"   
## [70] "-Dave Collins" "-Dave Concepcion" "-Darren Daulton"   
## [73] "-Doug DeCinces" "-Darrell Evans" "-Dwight Evans"   
## [76] "-Damaso Garcia" "-Dan Gladden" "-Danny Heep"   
## [79] "-Dave Henderson" "-Donnie Hill" "-Dave Kingman"   
## [82] "-Davey Lopes" "-Don Mattingly" "-Darryl Motley"   
## [85] "-Dale Murphy" "-Dwayne Murphy" "-Dave Parker"   
## [88] "-Dan Pasqua" "-Darrell Porter" "-Dick Schofield"   
## [91] "-Don Slaught" "-Darryl Strawberry" "-Dale Sveum"   
## [94] "-Danny Tartabull" "-Dickie Thon" "-Denny Walling"   
## [97] "-Dave Winfield" "-Enos Cabell" "-Eric Davis"   
## [100] "-Eddie Milner" "-Eddie Murray" "-Ernest Riles"   
## [103] "-Ed Romero" "-Ernie Whitt" "-Fred Lynn"   
## [106] "-Floyd Rayford" "-Franklin Stubbs" "-Frank White"   
## [109] "-George Bell" "-Glenn Braggs" "-George Brett"   
## [112] "-Greg Brock" "-Gary Carter" "-Glenn Davis"   
## [115] "-George Foster" "-Gary Gaetti" "-Greg Gagne"   
## [118] "-George Hendrick" "-Glenn Hubbard" "-Garth Iorg"   
## [121] "-Gary Matthews" "-Graig Nettles" "-Gary Pettis"   
## [124] "-Gary Redus" "-Garry Templeton" "-Gorman Thomas"   
## [127] "-Greg Walker" "-Gary Ward" "-Glenn Wilson"   
## [130] "-Harold Baines" "-Hubie Brooks" "-Howard Johnson"   
## [133] "-Hal McRae" "-Harold Reynolds" "-Harry Spilman"   
## [136] "-Herm Winningham" "-Jesse Barfield" "-Juan Beniquez"   
## [139] "-Juan Bonilla" "-John Cangelosi" "-Jose Canseco"   
## [142] "-Joe Carter" "-Jack Clark" "-Jose Cruz"   
## [145] "-Julio Cruz" "-Jody Davis" "-Jim Dwyer"   
## [148] "-Julio Franco" "-Jim Gantner" "-Johnny Grubb"   
## [151] "-Jerry Hairston" "-Jack Howell" "-John Kruk"   
## [154] "-Jeffrey Leonard" "-Jim Morrison" "-John Moses"   
## [157] "-Jerry Mumphrey" "-Joe Orsulak" "-Jorge Orta"   
## [160] "-Jim Presley" "-Jamie Quirk" "-Johnny Ray"   
## [163] "-Jeff Reed" "-Jim Rice" "-Jerry Royster"   
## [166] "-John Russell" "-Juan Samuel" "-John Shelby"   
## [169] "-Joel Skinner" "-Jeff Stone" "-Jim Sundberg"   
## [172] "-Jim Traber" "-Jose Uribe" "-Jerry Willard"   
## [175] "-Joel Youngblood" "-Kevin Bass" "-Kal Daniels"   
## [178] "-Kirk Gibson" "-Ken Griffey" "-Keith Hernandez"   
## [181] "-Kent Hrbek" "-Ken Landreaux" "-Kevin McReynolds"   
## [184] "-Kevin Mitchell" "-Keith Moreland" "-Ken Oberkfell"   
## [187] "-Ken Phelps" "-Kirby Puckett" "-Kurt Stillwell"   
## [190] "-Leon Durham" "-Len Dykstra" "-Larry Herndon"   
## [193] "-Lee Lacy" "-Len Matuszek" "-Lloyd Moseby"   
## [196] "-Lance Parrish" "-Larry Parrish" "-Luis Rivera"   
## [199] "-Larry Sheets" "-Lonnie Smith" "-Lou Whitaker"   
## [202] "-Mike Aldrete" "-Marty Barrett" "-Mike Brown"   
## [205] "-Mike Davis" "-Mike Diaz" "-Mariano Duncan"   
## [208] "-Mike Easler" "-Mike Fitzgerald" "-Mel Hall"   
## [211] "-Mickey Hatcher" "-Mike Heath" "-Mike Kingery"   
## [214] "-Mike LaValliere" "-Mike Marshall" "-Mike Pagliarulo"   
## [217] "-Mark Salas" "-Mike Schmidt" "-Mike Scioscia"   
## [220] "-Mickey Tettleton" "-Milt Thompson" "-Mitch Webster"   
## [223] "-Mookie Wilson" "-Marvell Wynne" "-Mike Young"   
## [226] "-Nick Esasky" "-Ozzie Guillen" "-Oddibe McDowell"   
## [229] "-Omar Moreno" "-Ozzie Smith" "-Ozzie Virgil"   
## [232] "-Phil Bradley" "-Phil Garner" "-Pete Incaviglia"   
## [235] "-Paul Molitor" "-Pete O'Brien" "-Pete Rose"   
## [238] "-Pat Sheridan" "-Pat Tabler" "-Rafael Belliard"   
## [241] "-Rick Burleson" "-Randy Bush" "-Rick Cerone"   
## [244] "-Ron Cey" "-Rob Deer" "-Rick Dempsey"   
## [247] "-Rich Gedman" "-Ron Hassey" "-Rickey Henderson"   
## [250] "-Reggie Jackson" "-Ricky Jones" "-Ron Kittle"   
## [253] "-Ray Knight" "-Randy Kutcher" "-Rudy Law"   
## [256] "-Rick Leach" "-Rick Manning" "-Rance Mulliniks"   
## [259] "-Ron Oester" "-Rey Quinones" "-Rafael Ramirez"   
## [262] "-Ronn Reynolds" "-Ron Roenicke" "-Ryne Sandberg"   
## [265] "-Rafael Santana" "-Rick Schu" "-Ruben Sierra"   
## [268] "-Roy Smalley" "-Robby Thompson" "-Rob Wilfong"   
## [271] "-Reggie Williams" "-Robin Yount" "-Steve Balboni"   
## [274] "-Scott Bradley" "-Sid Bream" "-Steve Buechele"   
## [277] "-Shawon Dunston" "-Scott Fletcher" "-Steve Garvey"   
## [280] "-Steve Jeltz" "-Steve Lombardozzi" "-Spike Owen"   
## [283] "-Steve Sax" "-Tony Armas" "-Tony Bernazard"   
## [286] "-Tom Brookens" "-Tom Brunansky" "-Tony Fernandez"   
## [289] "-Tim Flannery" "-Tom Foley" "-Tony Gwynn"   
## [292] "-Terry Harper" "-Toby Harrah" "-Tommy Herr"   
## [295] "-Tim Hulett" "-Terry Kennedy" "-Tito Landrum"   
## [298] "-Tim Laudner" "-Tom O'Malley" "-Tom Paciorek"   
## [301] "-Tony Pena" "-Terry Pendleton" "-Tony Perez"   
## [304] "-Tony Phillips" "-Terry Puhl" "-Tim Raines"   
## [307] "-Ted Simmons" "-Tim Teufel" "-Tim Wallach"   
## [310] "-Vince Coleman" "-Von Hayes" "-Vance Law"   
## [313] "-Wally Backman" "-Wade Boggs" "-Will Clark"   
## [316] "-Wally Joyner" "-Wayne Krenchicki" "-Willie McGee"   
## [319] "-Willie Randolph" "-Wayne Tolleson" "-Willie Upshaw"   
## [322] "-Willie Wilson"

# write the dataframe Hitters to a file  
write.table(Hitters, file = "Hitters.csv", row.names = TRUE)  
  
# check how many missing values there are in the Salary column of Hitters  
sum(is.na(Hitters$Salary))

## [1] 59

# remove missing values from Hitters and assign it to a dataframe called Hitters2  
Hitters2 = na.omit(Hitters)  
  
# check the dimension of Hitters2 (how many rows, how many columns)  
dim(Hitters2)

## [1] 263 20

# check how many missing values there are in the dataframe Hitters  
sum(is.na(Hitters2))

## [1] 0

## TASK 2 ##  
  
# load library leaps, which has the regsubsets() command for subset selection  
library(leaps)  
  
# run regsubsets using Salary as the repsonse, all other variables as predictors, with data from Hitters2  
# store the results in a variable called regfit.full  
regfit.full = regsubsets(Salary ~ ., data = Hitters2)  
summary(regfit.full)

## Subset selection object  
## Call: regsubsets.formula(Salary ~ ., data = Hitters2)  
## 19 Variables (and intercept)  
## Forced in Forced out  
## AtBat FALSE FALSE  
## Hits FALSE FALSE  
## HmRun FALSE FALSE  
## Runs FALSE FALSE  
## RBI FALSE FALSE  
## Walks FALSE FALSE  
## Years FALSE FALSE  
## CAtBat FALSE FALSE  
## CHits FALSE FALSE  
## CHmRun FALSE FALSE  
## CRuns FALSE FALSE  
## CRBI FALSE FALSE  
## CWalks FALSE FALSE  
## LeagueN FALSE FALSE  
## DivisionW FALSE FALSE  
## PutOuts FALSE FALSE  
## Assists FALSE FALSE  
## Errors FALSE FALSE  
## NewLeagueN FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI  
## 1 ( 1 ) " " " " " " " " " " " " " " " " " " " " " " "\*"   
## 2 ( 1 ) " " "\*" " " " " " " " " " " " " " " " " " " "\*"   
## 3 ( 1 ) " " "\*" " " " " " " " " " " " " " " " " " " "\*"   
## 4 ( 1 ) " " "\*" " " " " " " " " " " " " " " " " " " "\*"   
## 5 ( 1 ) "\*" "\*" " " " " " " " " " " " " " " " " " " "\*"   
## 6 ( 1 ) "\*" "\*" " " " " " " "\*" " " " " " " " " " " "\*"   
## 7 ( 1 ) " " "\*" " " " " " " "\*" " " "\*" "\*" "\*" " " " "   
## 8 ( 1 ) "\*" "\*" " " " " " " "\*" " " " " " " "\*" "\*" " "   
## CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN  
## 1 ( 1 ) " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " "\*" " " " " " "   
## 4 ( 1 ) " " " " "\*" "\*" " " " " " "   
## 5 ( 1 ) " " " " "\*" "\*" " " " " " "   
## 6 ( 1 ) " " " " "\*" "\*" " " " " " "   
## 7 ( 1 ) " " " " "\*" "\*" " " " " " "   
## 8 ( 1 ) "\*" " " "\*" "\*" " " " " " "

## TASK 3 ##  
  
# if you wanted to see best subsets with up to 15 predictors  
regfit.full = regsubsets(Salary ~ ., data = Hitters2, nvmax = 15)  
summary(regfit.full)

## Subset selection object  
## Call: regsubsets.formula(Salary ~ ., data = Hitters2, nvmax = 15)  
## 19 Variables (and intercept)  
## Forced in Forced out  
## AtBat FALSE FALSE  
## Hits FALSE FALSE  
## HmRun FALSE FALSE  
## Runs FALSE FALSE  
## RBI FALSE FALSE  
## Walks FALSE FALSE  
## Years FALSE FALSE  
## CAtBat FALSE FALSE  
## CHits FALSE FALSE  
## CHmRun FALSE FALSE  
## CRuns FALSE FALSE  
## CRBI FALSE FALSE  
## CWalks FALSE FALSE  
## LeagueN FALSE FALSE  
## DivisionW FALSE FALSE  
## PutOuts FALSE FALSE  
## Assists FALSE FALSE  
## Errors FALSE FALSE  
## NewLeagueN FALSE FALSE  
## 1 subsets of each size up to 15  
## Selection Algorithm: exhaustive  
## AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI  
## 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 ) "\*" "\*" "\*" "\*" " " "\*" " " "\*" "\*" " " "\*" "\*"   
## CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN  
## 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 ) "\*" "\*" "\*" "\*" "\*" "\*" " "

## TASK 4 ##  
  
# store the sumary of results in a variable called reg.summary  
reg.summary = summary(regfit.full)  
  
# check the names of columns in reg.summary  
names(reg.summary)

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

# display the column reg.summary$rsq  
# what you see in the output is R^2 for the best subset with one predictor,  
# R^2 for the best subset with two predictors, etc.  
reg.summary$rsq

## [1] 0.3214501 0.4252237 0.4514294 0.4754067 0.4908036 0.5087146 0.5141227  
## [8] 0.5285569 0.5346124 0.5404950 0.5426153 0.5436302 0.5444570 0.5452164  
## [15] 0.5454692

# display the column reg.summary$adjr2  
# what you see in the output is adjusted R^2 for the best subset with one predictor,  
# adjusted R^2 for the best subset with two predictors, etc.  
reg.summary$adjr2

## [1] 0.3188503 0.4208024 0.4450753 0.4672734 0.4808971 0.4972001 0.5007849  
## [8] 0.5137083 0.5180572 0.5222606 0.5225706 0.5217245 0.5206736 0.5195431  
## [15] 0.5178661

# display the column reg.summary$cp  
# what you see in the output is Cp for the best subset with one predictor,  
# Cp for the best subset with two predictors, etc.  
reg.summary$cp

## [1] 104.281319 50.723090 38.693127 27.856220 21.613011 14.023870  
## [7] 13.128474 7.400719 6.158685 5.009317 5.874113 7.330766  
## [13] 8.888112 10.481576 12.346193

which.max(reg.summary$adjr2) # function which.max() returns the index of maximum value

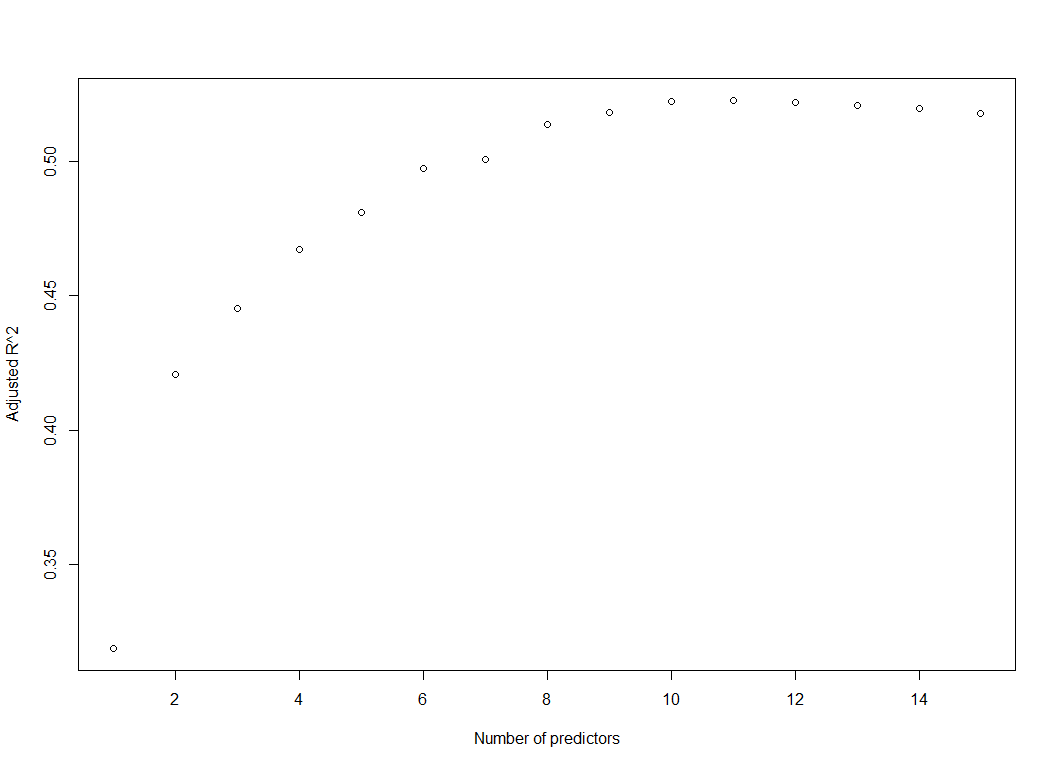
## [1] 11

coef(regfit.full, 11)

## (Intercept) AtBat Hits Walks CAtBat CRuns   
## 135.7512195 -2.1277482 6.9236994 5.6202755 -0.1389914 1.4553310   
## CRBI CWalks LeagueN DivisionW PutOuts Assists   
## 0.7852528 -0.8228559 43.1116152 -111.1460252 0.2894087 0.2688277

## TASK 5 ##  
  
# plot adjusted R^2 as a function of predictors in the model

plot(reg.summary$adjr2, xlab = "Number of predictors", ylab = "Adjusted R^2")



# display the number of predictors for which adjusted R^2 reaches its maximum -- it is 11  
which.max(reg.summary$adjr2)

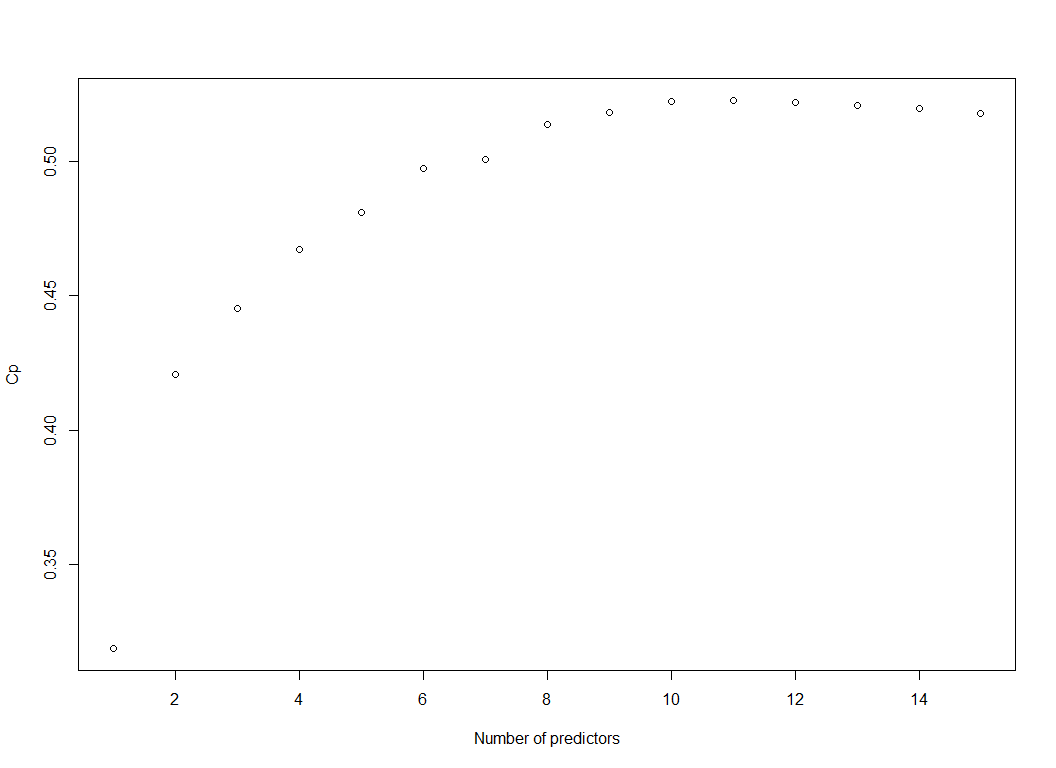
## [1] 11

# display the coefficient estimates for the best model with 11 predictors  
coef(regfit.full, 11)

## (Intercept) AtBat Hits Walks CAtBat CRuns   
## 135.7512195 -2.1277482 6.9236994 5.6202755 -0.1389914 1.4553310   
## CRBI CWalks LeagueN DivisionW PutOuts Assists   
## 0.7852528 -0.8228559 43.1116152 -111.1460252 0.2894087 0.2688277

## TASK 6 ##  
  
# plot Cp as a function of predictors in the model

plot(reg.summary$adjr2, xlab = "Number of predictors", ylab = "Cp")



# display the number of predictors for which Cp reaches its minimum -- it is 10  
which.min(reg.summary$cp)

## [1] 10

# display the coefficient estimates for the best model with 10 predictors  
coef(regfit.full, 10)

## (Intercept) AtBat Hits Walks CAtBat CRuns   
## 162.5354420 -2.1686501 6.9180175 5.7732246 -0.1300798 1.4082490   
## CRBI CWalks DivisionW PutOuts Assists   
## 0.7743122 -0.8308264 -112.3800575 0.2973726 0.2831680

## TASK 7 ##  
  
# run forward stepwise selection, allowing subsets with up to 19 predictors  
regfit.fwd = regsubsets(Salary ~ .,  
 data = Hitters2,  
 nvmax = 19,  
 method = "forward")  
summary(regfit.fwd)

## Subset selection object  
## Call: regsubsets.formula(Salary ~ ., data = Hitters2, nvmax = 19, method = "forward")  
## 19 Variables (and intercept)  
## Forced in Forced out  
## AtBat FALSE FALSE  
## Hits FALSE FALSE  
## HmRun FALSE FALSE  
## Runs FALSE FALSE  
## RBI FALSE FALSE  
## Walks FALSE FALSE  
## Years FALSE FALSE  
## CAtBat FALSE FALSE  
## CHits FALSE FALSE  
## CHmRun FALSE FALSE  
## CRuns FALSE FALSE  
## CRBI FALSE FALSE  
## CWalks FALSE FALSE  
## LeagueN FALSE FALSE  
## DivisionW FALSE FALSE  
## PutOuts FALSE FALSE  
## Assists FALSE FALSE  
## Errors FALSE FALSE  
## NewLeagueN FALSE FALSE  
## 1 subsets of each size up to 19  
## Selection Algorithm: forward  
## AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI  
## 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 ) "\*" "\*" "\*" "\*" "\*" "\*" " " "\*" "\*" " " "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN  
## 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 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

## TASK 8 ##  
  
# run backward stepwise selection, allowing subsets with up to 19 predictors  
regfit.bwd = regsubsets(Salary ~ .,  
 data = Hitters2,  
 nvmax = 19,  
 method = "backward")  
summary(regfit.bwd)

## Subset selection object  
## Call: regsubsets.formula(Salary ~ ., data = Hitters2, nvmax = 19, method = "backward")  
## 19 Variables (and intercept)  
## Forced in Forced out  
## AtBat FALSE FALSE  
## Hits FALSE FALSE  
## HmRun FALSE FALSE  
## Runs FALSE FALSE  
## RBI FALSE FALSE  
## Walks FALSE FALSE  
## Years FALSE FALSE  
## CAtBat FALSE FALSE  
## CHits FALSE FALSE  
## CHmRun FALSE FALSE  
## CRuns FALSE FALSE  
## CRBI FALSE FALSE  
## CWalks FALSE FALSE  
## LeagueN FALSE FALSE  
## DivisionW FALSE FALSE  
## PutOuts FALSE FALSE  
## Assists FALSE FALSE  
## Errors FALSE FALSE  
## NewLeagueN FALSE FALSE  
## 1 subsets of each size up to 19  
## Selection Algorithm: backward  
## AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI  
## 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 ) "\*" "\*" "\*" "\*" "\*" "\*" " " "\*" "\*" " " "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN  
## 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 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

## TASK 9 ##  
  
# the coefficient estimates for the best model with 7 predictors when using best subset selection  
coef(regfit.full, 7)

## (Intercept) Hits Walks CAtBat CHits CHmRun   
## 79.4509472 1.2833513 3.2274264 -0.3752350 1.4957073 1.4420538   
## DivisionW PutOuts   
## -129.9866432 0.2366813

# the coefficient estimates for the best model with 7 predictors when using forward selection  
coef(regfit.fwd, 7)

## (Intercept) AtBat Hits Walks CRBI CWalks   
## 109.7873062 -1.9588851 7.4498772 4.9131401 0.8537622 -0.3053070   
## DivisionW PutOuts   
## -127.1223928 0.2533404

# the coefficient estimates for the best model with 7 predictors when using backward selection  
coef(regfit.bwd, 7)

## (Intercept) AtBat Hits Walks CRuns CWalks   
## 105.6487488 -1.9762838 6.7574914 6.0558691 1.1293095 -0.7163346   
## DivisionW PutOuts   
## -116.1692169 0.3028847

###############################################  
#### in-class exercise  
library(leaps)  
Auto = read.csv(  
 "Auto.csv",  
 header = T,  
 na.strings = "?",  
 stringsAsFactors = T  
)  
Auto2 = na.omit(Auto)  
  
head(Auto2)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

my.auto = regsubsets(mpg ~ . - name, data = Auto2)  
summary(my.auto)

## Subset selection object  
## Call: regsubsets.formula(mpg ~ . - name, data = Auto2)  
## 7 Variables (and intercept)  
## Forced in Forced out  
## cylinders FALSE FALSE  
## displacement FALSE FALSE  
## horsepower FALSE FALSE  
## weight FALSE FALSE  
## acceleration FALSE FALSE  
## year FALSE FALSE  
## origin FALSE FALSE  
## 1 subsets of each size up to 7  
## Selection Algorithm: exhaustive  
## cylinders displacement horsepower weight acceleration year origin  
## 1 ( 1 ) " " " " " " "\*" " " " " " "   
## 2 ( 1 ) " " " " " " "\*" " " "\*" " "   
## 3 ( 1 ) " " " " " " "\*" " " "\*" "\*"   
## 4 ( 1 ) " " "\*" " " "\*" " " "\*" "\*"   
## 5 ( 1 ) " " "\*" "\*" "\*" " " "\*" "\*"   
## 6 ( 1 ) "\*" "\*" "\*" "\*" " " "\*" "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

coef(my.auto, 3)

## (Intercept) weight year origin   
## -18.045850149 -0.005994118 0.757126111 1.150390789

my.auto2 = regsubsets(mpg ~ . - name, data = Auto2, method = "forward")  
coef(my.auto2, 3)

## (Intercept) weight year origin   
## -18.045850149 -0.005994118 0.757126111 1.150390789

my.auto3 = regsubsets(mpg ~ . - name, data = Auto2, method = "backward")  
coef(my.auto3, 3)

## (Intercept) weight year origin   
## -18.045850149 -0.005994118 0.757126111 1.150390789

#################################################