Project_2

Lee Panter, Arlin Tawzer, Nick Weaver

10/15/2018

Markdown and Knitr options

Working directories

```
P2WD="/Users/lee/Desktop/MATH_6388/Project_2"
setwd(P2WD)
```

Libraries & Packages:

- lubridate

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(leaps)
library(car)
```

Data Dependencies:

- nonrookies.rda – Note: this is a pre-altered dataset that has had rookies and several variables merged and removed.

```
load(file = "/Users/lee/Desktop/MATH_6388/Project_2/nonrookies.rda")
dat=nonrookies
rm(nonrookies)
```

Initial Data Analysis

```
(names=names(dat))
```

```
##  [1] "index"          "yearID"         "teamID"         "lgID"
##  [5] "playerID"       "salary_2015"    "salary_2016"    "birthYear"
##  [9] "birthMonth"     "birthDay"       "birthCountry"   "birthState"
## [13] "birthCity"      "deathYear"      "deathMonth"     "deathDay"
## [17] "deathCountry"   "deathState"     "deathCity"      "nameFirst"
## [21] "nameLast"       "nameGiven"      "weight"         "height"
## [25] "bats"           "throws"         "debut"          "finalGame"
## [29] "retroID"        "bbrefID"        "G"              "AB"
## [33] "R"              "H"              "X2B"            "X3B"
## [37] "HR"             "RBI"            "SB"             "CS"
## [41] "BB"             "SO"             "IBB"            "HBP"
## [45] "SH"             "SF"             "GIDP"           "debut_year"
## [49] "finalgame_year"
```

```
summary(dat)
```

```
##      index           yearID         teamID    lgID          playerID
##  Min.   :24761   Min.   :2015   BOS    : 28   AL:370   abadfe01 :  1
##  1st Qu.:24971   1st Qu.:2015   MIL    : 28   NL:355   ackledu01:  1
##  Median :25168   Median :2015   BAL    : 27            adamsma01:  1
##  Mean   :25168   Mean   :2015   PIT    : 27            affelje01:  1
##  3rd Qu.:25369   3rd Qu.:2015   HOU    : 26            alberma01:  1
##  Max.   :25574   Max.   :2015   KCA    : 26            albural01:  1
##                                 (Other):563            (Other)  :719
##   salary_2015        salary_2016         birthYear      birthMonth
##  Min.   :  507000   Min.   :  507500   Min.   :1972   Min.   : 1.000
##  1st Qu.:  543000   1st Qu.: 1400000   1st Qu.:1983   1st Qu.: 4.000
##  Median : 2350000   Median : 4000000   Median :1985   Median : 7.000
##  Mean   : 4695632   Mean   : 6289413   Mean   :1985   Mean   : 6.532
##  3rd Qu.: 6666000   3rd Qu.: 8614388   3rd Qu.:1988   3rd Qu.: 9.000
##  Max.   :32571000   Max.   :33000000   Max.   :1992   Max.   :12.000
##                     NA's   :178
##     birthDay        birthCountry             birthState
##  Min.   : 1.00   USA      :531   CA               :111
##  1st Qu.: 8.00   D.R.     : 70   TX               : 57
##  Median :16.00   Venezuela: 56   FL               : 55
##  Mean   :15.72   Cuba     : 13   IL               : 25
##  3rd Qu.:22.00   P.R.     : 11   Distrito Nacional: 24
##  Max.   :31.00   Mexico   :  8   GA               : 23
##                  (Other)  : 36   (Other)          :430
##          birthCity     deathYear      deathMonth     deathDay
##  Santo Domingo: 23   Min.   :2016   Min.   :1     Min.   :22.00
##  Houston      : 16   1st Qu.:2016   1st Qu.:3     1st Qu.:22.75
##  Valencia     : 10   Median :2016   Median :5     Median :23.50
##  Atlanta      :  7   Mean   :2016   Mean   :5     Mean   :23.50
##  San Diego    :  7   3rd Qu.:2017   3rd Qu.:7     3rd Qu.:24.25
##  Santiago     :  7   Max.   :2017   Max.   :9     Max.   :25.00
##  (Other)      :655   NA's   :723    NA's   :723   NA's   :723
##  deathCountry          deathState        deathCity     nameFirst
##      :723                   :723             :723   Chris  : 17
##  D.R.:  1     FL            :  1   Juan Adrian:  1   Matt   : 15
##  USA :  1     Monsenor Nouel:  1   Miami Beach:  1   David  : 12
##                                                      Justin : 11
##                                                      Carlos : 10
##                                                      Jason  : 10
##                                                      (Other):650
##       nameLast             nameGiven       weight          height
##  Rodriguez:  6   Anthony Michael:  4   Min.   :160.0   Min.   :66.00
##  Davis    :  5   James Anthony  :  3   1st Qu.:200.0   1st Qu.:72.00
##  Gonzalez :  5   Jason Michael  :  3   Median :210.0   Median :74.00
##  Ramirez  :  5   Matthew Thomas :  3   Mean   :213.2   Mean   :73.65
##  Cabrera  :  4   Adam Parrish   :  2   3rd Qu.:225.0   3rd Qu.:75.00
##  Hernandez:  4   Alberto Jose   :  2   Max.   :300.0   Max.   :82.00
##  (Other)  :696   (Other)        :708
##  bats    throws       debut         finalGame       retroID
##  B: 63   L:162   9/1/2010:  6   10/1/2017:198   abadf001:  1
##  L:219   R:563   4/1/2013:  4   9/30/2017: 96   ackld001:  1
##  R:443           4/2/2007:  4   9/29/2017: 43   adamm002:  1
##                  4/3/2006:  4   9/28/2017: 25   affej001:  1
##                  9/2/2008:  4   10/2/2016: 23   albem001:  1
```

```
##                 9/2/2011:  4   10/4/2015: 18   albua001:  1
##                 (Other) :699   (Other)  :322   (Other) :719
##       bbrefID          G                AB              R
##  abadfe01 :  1   Min.   :  11.0   Min.   :    0   Min.   :   0.0
##  ackledu01:  1   1st Qu.: 249.0   1st Qu.:   15   1st Qu.:   1.0
##  adamsma01:  1   Median : 448.0   Median :  493   Median :  39.0
##  affelje01:  1   Mean   : 607.6   Mean   : 1622   Mean   : 219.3
##  alberma01:  1   3rd Qu.: 831.0   3rd Qu.: 2745   3rd Qu.: 354.0
##  albural01:  1   Max.   :2814.0   Max.   :10635   Max.   :2021.0
##  (Other)  :719
##        H              X2B              X3B               HR
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:   2.0   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
##  Median :  89.0   Median : 18.00   Median :  1.000   Median :  6.00
##  Mean   : 429.1   Mean   : 85.56   Mean   :  9.393   Mean   : 52.46
##  3rd Qu.: 705.0   3rd Qu.:139.00   3rd Qu.: 14.000   3rd Qu.: 73.00
##  Max.   :3115.0   Max.   :632.00   Max.   :128.000   Max.   :696.00
##
##       RBI              SB               CS               BB
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0.0
##  Median :  34.0   Median :  1.00   Median :  1.00   Median :  25.0
##  Mean   : 208.5   Mean   : 31.18   Mean   : 11.14   Mean   : 150.6
##  3rd Qu.: 316.0   3rd Qu.: 30.00   3rd Qu.: 15.00   3rd Qu.: 235.0
##  Max.   :2086.0   Max.   :512.00   Max.   :125.00   Max.   :1338.0
##
##        SO              IBB              HBP               SH
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:   7.0   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 155.0   Median :  0.00   Median :  2.00   Median :  4.00
##  Mean   : 341.1   Mean   : 12.13   Mean   : 15.97   Mean   : 11.61
##  3rd Qu.: 568.0   3rd Qu.: 13.00   3rd Qu.: 23.00   3rd Qu.: 17.00
##  Max.   :2287.0   Max.   :307.00   Max.   :199.00   Max.   :100.00
##
##        SF             GIDP          debut_year   finalgame_year
##  Min.   :  0.0   Min.   :  0.00   Min.   :1994   Min.   :2002
##  1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.:2006   1st Qu.:2016
##  Median :  2.0   Median :  8.00   Median :2009   Median :2017
##  Mean   : 12.8   Mean   : 36.43   Mean   :2009   Mean   :2017
##  3rd Qu.: 19.0   3rd Qu.: 55.00   3rd Qu.:2011   3rd Qu.:2017
##  Max.   :111.0   Max.   :362.00   Max.   :2013   Max.   :2017
##
```

Removal of non-quantitative predictors

```
dat <- subset(dat, select = -c(yearID,teamID,lgID,playerID,
                               birthYear,birthMonth,birthDay,
                               birthCountry,birthState,birthCity,
                               deathYear,deathMonth,deathDay,
                               deathCountry,deathState,deathCity,
                               nameFirst,nameLast,nameGiven,bats,
                               throws,debut,finalGame,retroID,
                               bbrefID,finalgame_year))
```

Set values of NA salary to 0, remove 0-salaried players

```
for (i in 1:nrow(dat)){
  temp.sal.16 <- dat$salary_2016[i]
  temp.sal.15 <- dat$salary_2015[i]
  if (is.na(temp.sal.16))
  {dat$salary_2016[i] = 0}
  else if (is.na(temp.sal.15))
  {dat$salary_2015[i]=0}
}
dat=subset(dat, (salary_2016 != 0) & (salary_2015 != 0), select = index:GIDP)

rm(i); rm(temp.sal.15); rm(temp.sal.16)
```

Split data into training and test sets

```
set.seed(123)
a <- 0.7*nrow(dat)
tmp.random<-sample(1:nrow(dat))
training<-dat[tmp.random[1:a],]
dim(training)
```

```
## [1] 382  22
```

```
test=dat[-tmp.random[1:a],]
dim(test)
```

```
## [1] 165  22
```

Create OLS model

```
lmod.train.15=lm(salary_2015~.-salary_2016, data = training)
lmod.train.15.s=summary(lmod.train.15)
```

regsubsets

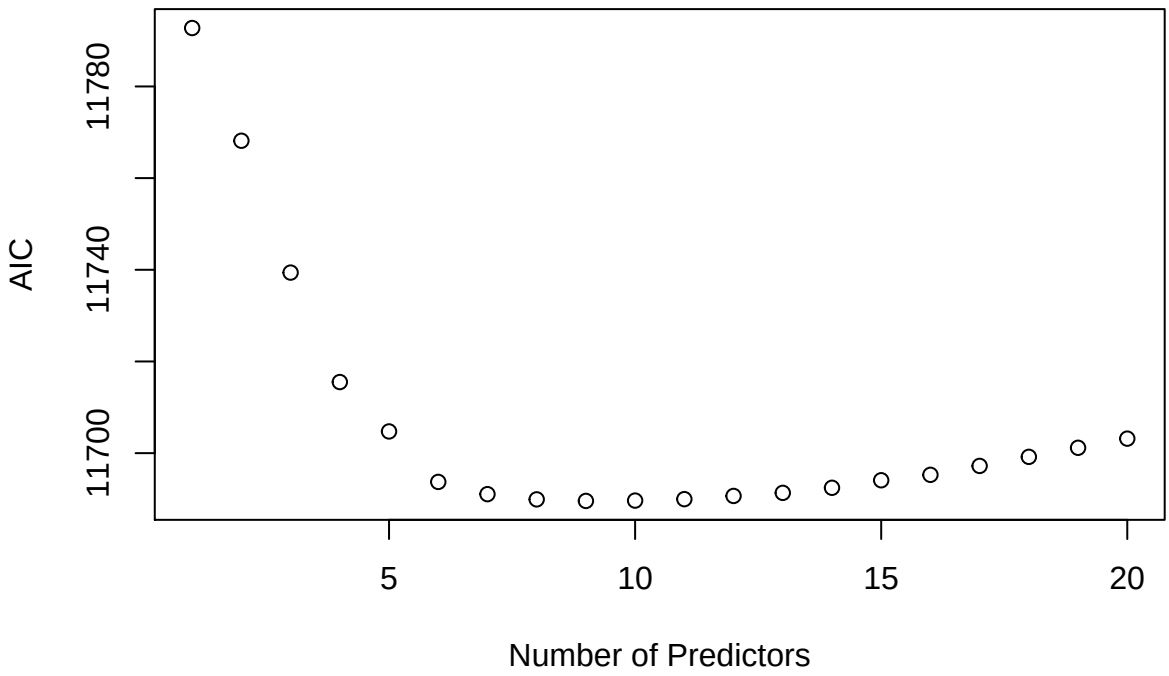
```
regfit.train.15=regsubsets(salary_2015~.-salary_2016, data = training, nvmax = 21)
regfit.train.15.s=summary(regfit.train.15)
regfit.train.15.s$which
```

```
##    (Intercept) index weight height     G    AB     R     H   X2B   X3B
## 1         TRUE FALSE  FALSE  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 2         TRUE FALSE  FALSE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3         TRUE FALSE  FALSE  FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 4         TRUE FALSE  FALSE  FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## 5         TRUE FALSE  FALSE   TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
## 6         TRUE FALSE   TRUE  FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 7         TRUE FALSE   TRUE  FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 8         TRUE FALSE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 9         TRUE FALSE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 10        TRUE FALSE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 11        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 12        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 13        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 14        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 15        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 16        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 17        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 18        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
## 19        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 20        TRUE  TRUE   TRUE   TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##       HR   RBI    SB    CS    BB    SO   IBB   HBP    SH    SF  GIDP
## 1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 4  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 5  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 6  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## 7  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## 8  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
## 9  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 10 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 11 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE
## 12 FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 13  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 14  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 15  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 16  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 17  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 18  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 19  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 20  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Define and Plot AIC

```
n.train=dim(training)[1]
AIC=n.train*log(regfit.train.15.s$rss/n.train)+(1:20)*2
plot(AIC ~ I(1:20), ylab="AIC", xlab="Number of Predictors")
```



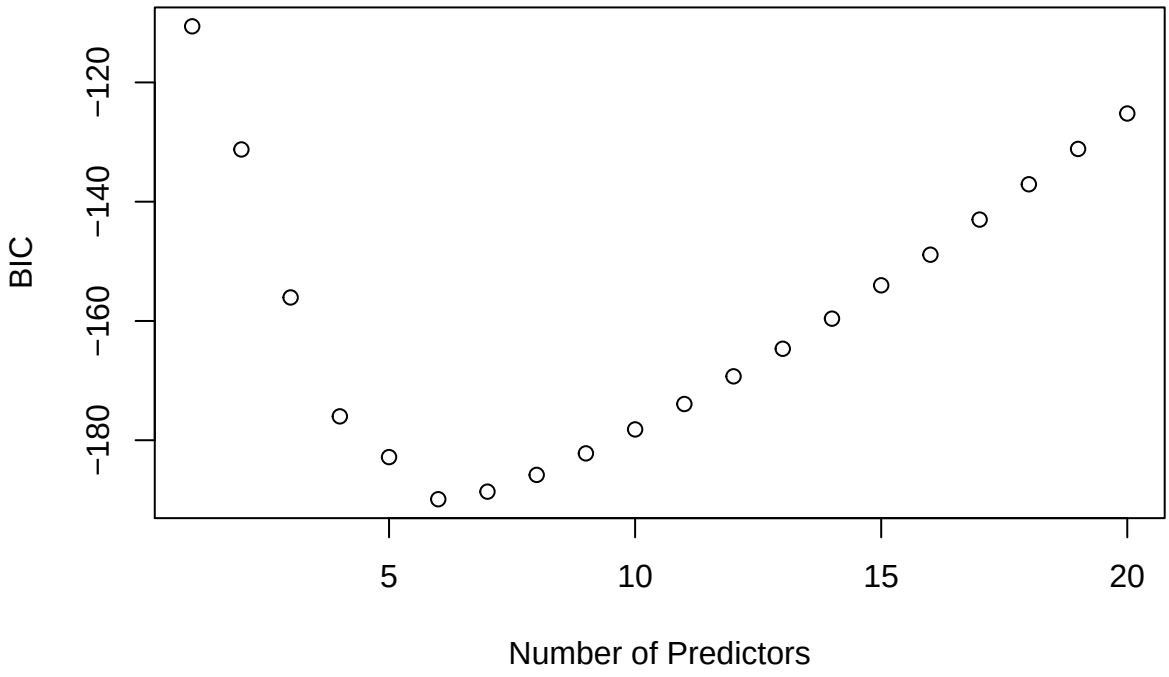

```
which.min(AIC)
```

```
## [1] 9
```

```
#The Number of predictors that minimizes AIC is 9
```

Define and Plot BIC

```
BIC=regfit.train.15.s$bic
plot(BIC ~ I(1:20), ylab="BIC", xlab="Number of Predictors")
```



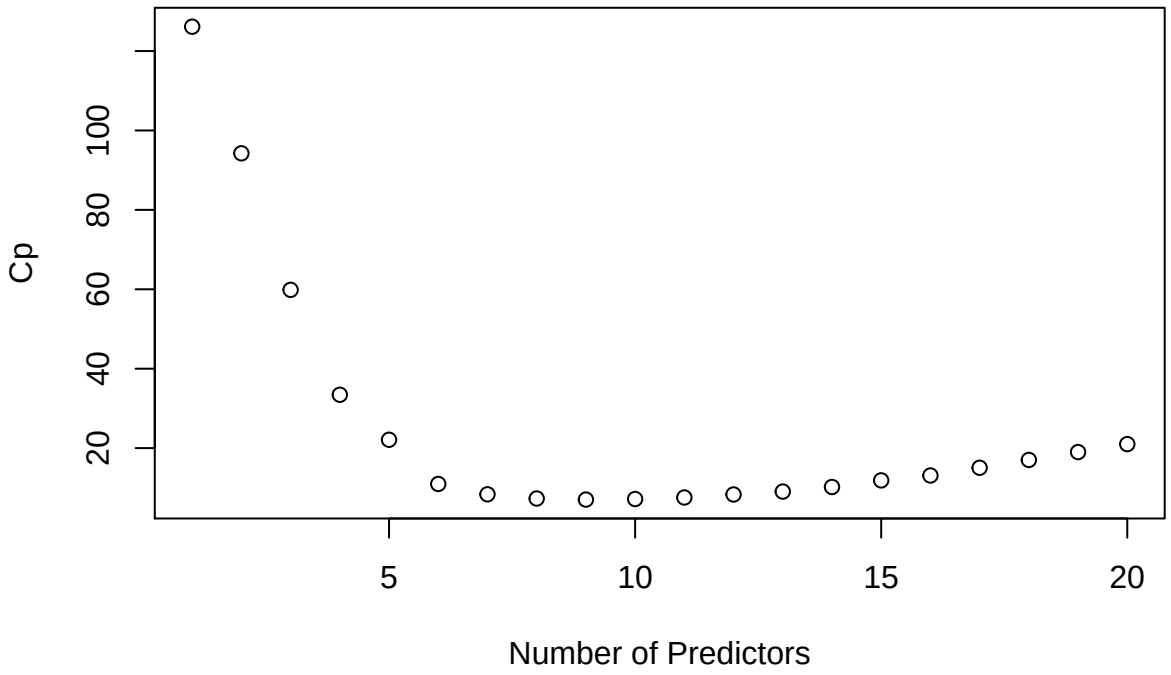

```
which.min(BIC)
```

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

```
#The number of predictors that minimizes BIC is 6
```

Define and Plot Cp

```
cp=regfit.train.15.s$cp
plot(cp~I(1:20), ylab="Cp", xlab="Number of Predictors")
```




```
which.min(cp)
```

```
## [1] 9
```

```
#The number of predictors that minimizes Cp is 9
```

Define New "Best" Functions according to the best 6 and 9 predictors

- Best six are:
- weight
- G
- AB
- H
- RBI
- SH
- Best nine are:
- weight
- height
- G
- AB
- H
- RBI
- HBP
- SH
- GIDP

```
lm.Acp=lm(salary_2015~weight+G+AB+H+RBI+SH, data = training)
lm.BIC=lm(salary_2015~weight+height+G+AB+H+RBI+HBP+SH+GIDP, data=training)
```

Calculate Predicted values for test set

```
pred.Acp.2015=predict.lm(lm.Acp, newdata = test)
pred.BIC.2015=predict.lm(lm.BIC, newdata = test)

SE.pred.Acp.2015=c()
SE.pred.BIC.2015=c()

for (i in 1:length(pred.Acp.2015))
{
  SE.pred.Acp.2015[i]=(test$salary_2015[i]-pred.Acp.2015[i])^2
  SE.pred.BIC.2015[i]=(test$salary_2015[i]-pred.BIC.2015[i])^2
}

(MSE.Acp.2015=sum(SE.pred.Acp.2015)/length(pred.Acp.2015))
```

```
## [1] 2.180882e+13
```

```
(MSE.BIC.2015=sum(SE.pred.BIC.2015)/length(pred.BIC.2015))
```

```
## [1] 2.133805e+13
```

```
#The smaller of the MSE measurements is for BIC (NOT that it is any good)
```

We now apply this model (BIC-selected variable model) towards prediction of the 2016 salaries, and calculate a loss function on this verification set:

```
dat2=dat[,-3]
dat2$salary_2015=dat$salary_2016
pred.BIC.2016=predict.lm(lm.BIC, newdata = dat2)

SE.pred.BIC.2016=c()

for (i in 1:length(pred.BIC.2016))
{
  SE.pred.BIC.2016[i]=(dat$salary_2016[i]-pred.BIC.2016[i])^2
}

(MSE.BIC.2016=sum(SE.pred.BIC.2016)/length(pred.BIC.2016))
```

```
## [1] 2.735433e+13
```