Lab2

Tristen Tooming

1/18/2021

1

The data set Hitters, which can be obtained from the ISLR library, contains data with 322 observations of major league baseball players on 20 variables that are assumed to influence the salary of the players.

The purpose of this assignment is to compare the results of

Best Subset Selection Forward Stepwise Selection Backward Stepwise Selection

Use the leaps library and set up models with all variables as predictors and Salary as response. Find and present the best models based on both BIC and Cp. Which method performs best based on the lowest BIC and Cp, and which variables are important for the salary for this model?

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

summary(Hitters)
```

```
     AtBat            Hits         HmRun            Runs
 Min.   : 16.0   Min.   :  1   Min.   : 0.00   Min.   :  0.00
 1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   1st Qu.: 30.25
 Median :379.5   Median : 96   Median : 8.00   Median : 48.00
 Mean   :380.9   Mean   :101   Mean   :10.77   Mean   : 50.91
 3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00   3rd Qu.: 69.00
 Max.   :687.0   Max.   :238   Max.   :40.00   Max.   :130.00

      RBI             Walks            Years            CAtBat
 Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
 1st Qu.: 28.00   1st Qu.: 22.00   1st Qu.: 4.000   1st Qu.:  816.8
 Median : 44.00   Median : 35.00   Median : 6.000   Median : 1928.0
 Mean   : 48.03   Mean   : 38.74   Mean   : 7.444   Mean   : 2648.7
 3rd Qu.: 64.75   3rd Qu.: 53.00   3rd Qu.:11.000   3rd Qu.: 3924.2
 Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0

     CHits            CHmRun           CRuns             CRBI
 Min.   :   4.0   Min.   :  0.00   Min.   :   1.0   Min.   :   0.00
 1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75
 Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
 Mean   : 717.6   Mean   : 69.49   Mean   : 358.8   Mean   : 330.12
 3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
 Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.00

     CWalks        League  Division    PutOuts          Assists
 Min.   :    0.00  A:175   E:157    Min.   :    0.0   Min.   :  0.0
 1st Qu.:   67.25  N:147   W:165    1st Qu.: 109.2   1st Qu.:  7.0
 Median : 170.50                    Median : 212.0   Median : 39.5
 Mean   : 260.24                    Mean   : 288.9   Mean   :106.9
 3rd Qu.: 339.25                    3rd Qu.: 325.0   3rd Qu.:166.0
 Max.   :1566.00                    Max.   :1378.0   Max.   :492.0

     Errors           Salary       NewLeague
 Min.   : 0.00   Min.   :  67.5   A:176
 1st Qu.: 3.00   1st Qu.: 190.0   N:146
 Median : 6.00   Median : 425.0
 Mean   : 8.04   Mean   : 535.9
 3rd Qu.:11.00   3rd Qu.: 750.0
 Max.   :32.00   Max.   :2460.0
                 NA's   :59
```

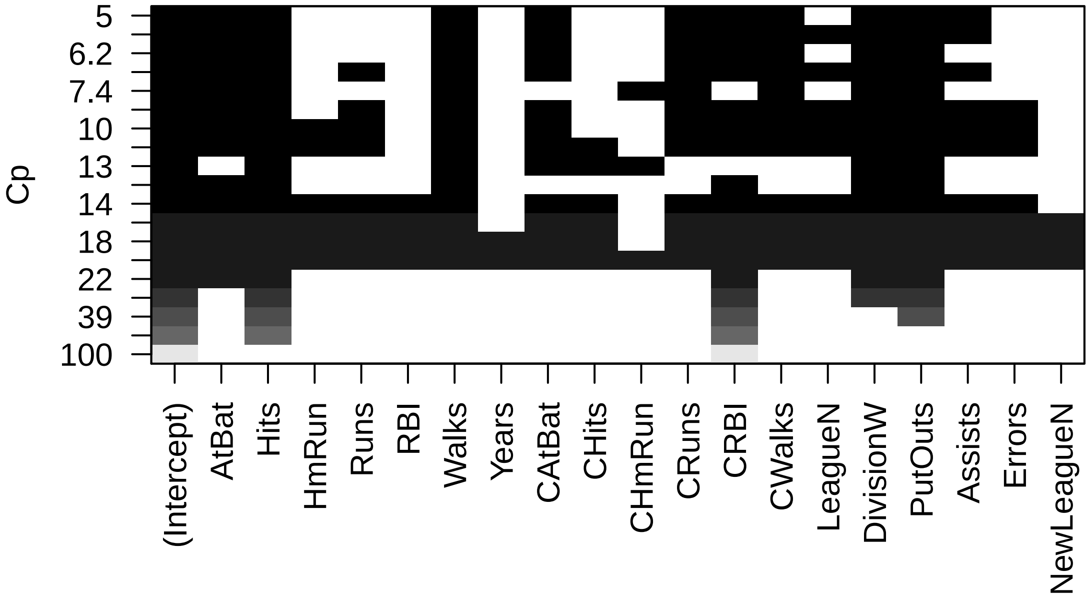
```r
# Fit
# Showing only best (nbest = 1)
regfit.models <- regsubsets(Salary~.,
                            data = Hitters,
                            nbest = 1,
                            nvmax = ncol(Hitters))

# Summary, Cp, BIC
res.sum <- summary(regfit.models)
as.data.frame(res.sum$outmat)
```

Best Subset Selection

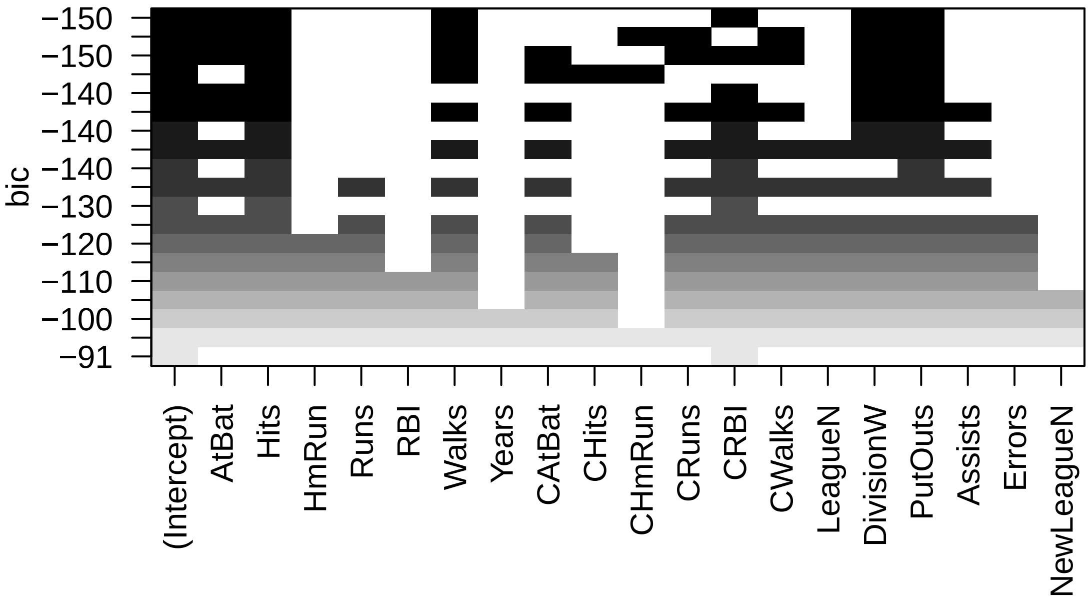
```
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
1  ( 1 )                                                                     *
2  ( 1 )           *                                                         *
3  ( 1 )           *                                                         *
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 )      *       *         *       *       *      *          *
```

```
plot(regfit.models, scale='Cp')
```



```
plot(regfit.models, scale='bic')
```



```
Best Cp model: 10

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CAtBat + CRuns + CRBI + CWalks + DivisionW + PutOu

Best BIC model: 6

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CRBI + DivisionW + PutOuts
```

```
regfit.fwd = regsubsets(Salary~.,
                        data=Hitters,
                        nvmax=ncol(Hitters),
                        method ="forward")

regfit.fwd.sum = summary(regfit.fwd)

regfit.fwd.sum.min.bic = which.min(regfit.fwd.sum$bic)
regfit.fwd.sum.min.cp = which.min(regfit.fwd.sum$cp)
```

Forward Stepwise Selection

```
Best Cp model: 10

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CAtBat + CRuns + CRBI + CWalks + DivisionW + PutOu

Best BIC model: 6

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CRBI + DivisionW + PutOuts
```

```
regfit.bfd = regsubsets(Salary~.,
                        data=Hitters,
                        nvmax=ncol(Hitters),
                        method ="backward")
regfit.bfd.sum = summary(regfit.bfd)
```

```r
regfit.bfd.sum.min.bic = which.min(regfit.bfd.sum$bic)
regfit.bfd.sum.min.cp = which.min(regfit.bfd.sum$cp)
```

Backward Stepwise Selection

```
Best Cp model: 10

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CAtBat + CRuns + CRBI + CWalks + DivisionW + PutOut

Best BIC model: 8

Model: Salary ~ (Intercept) + AtBat + Hits + Walks + CRuns + CRBI + CWalks + DivisionW + PutOuts
```