17) Subset Selection

Vitor Kamada

February 2018

Reference

Tables, Graphics, and Figures from

An Introduction to Statistical Learning

James et al. (2017): Chapters: 6.1, and 6.5

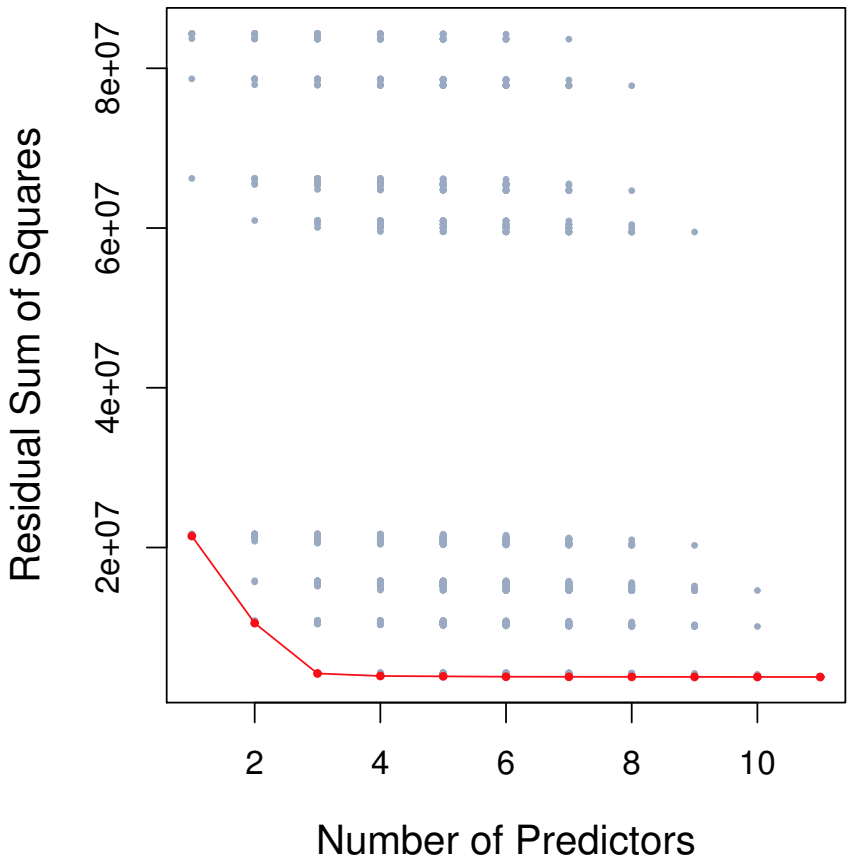
Best Subset Selection

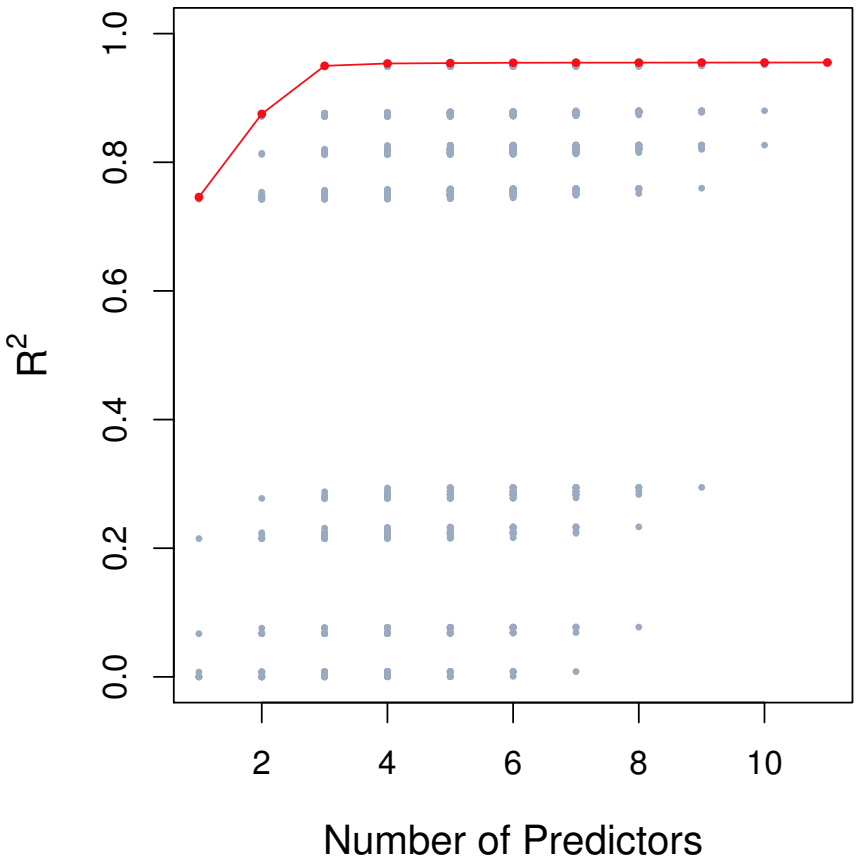
1) Let $\mathbb{M}_0$ denote the null model

2) Fit all $\binom{p}{k}$ models, and pick the best for each $\mathbb{M}_k$

3) Pick the single best among $\mathbb{M}_0, \ldots \mathbb{M}_p$ using cross-validated prediction error, C_p, AIC, BIC, or ajusted R^2

Credit Data Set

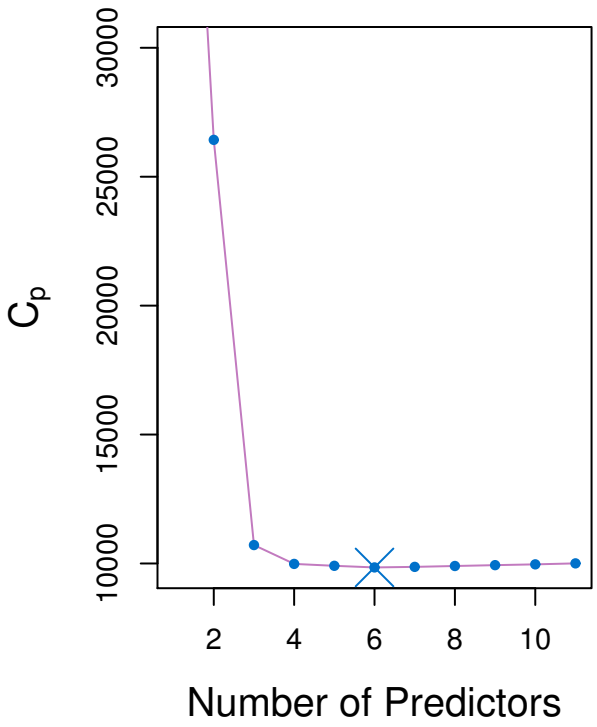


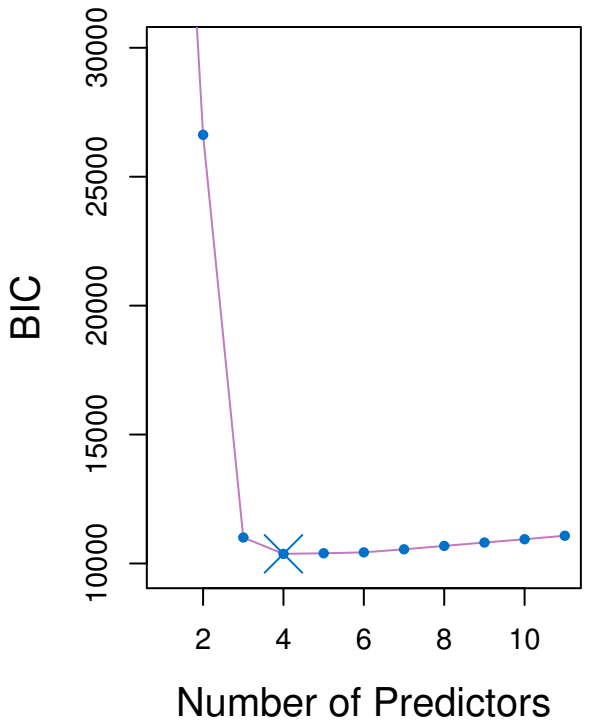


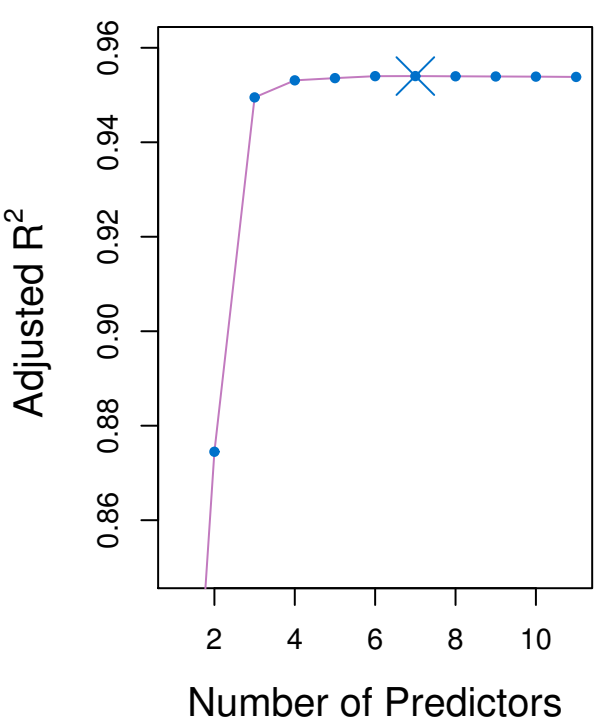
Best Subset vs Forward Stepwise

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

Test Errot: Adjusting the Training Error







C_p Statistic

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

$$\hat{\sigma}^2 = Var(\epsilon)$$

$$d = \# \text{ of predictors}$$

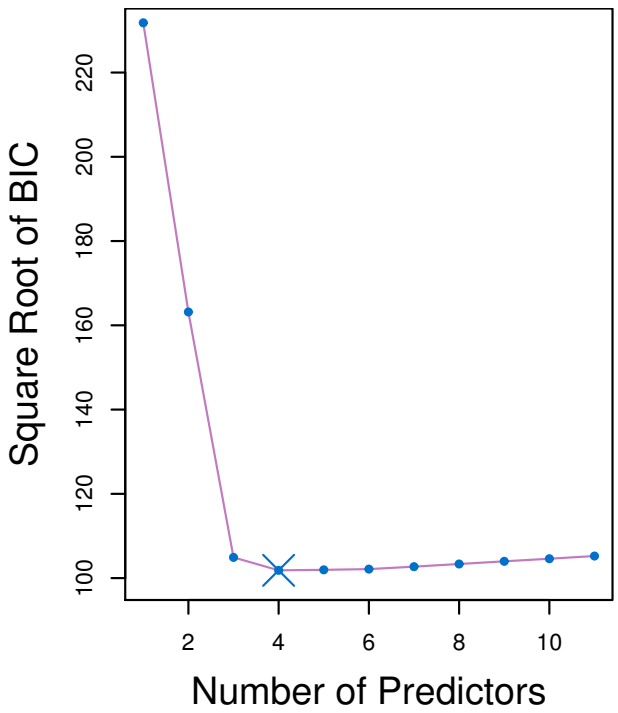
Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Adjusted R^2

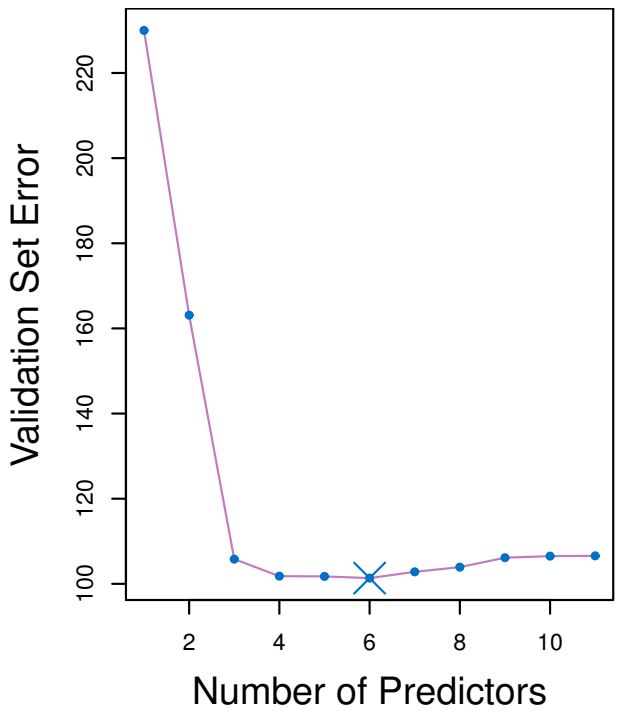
$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

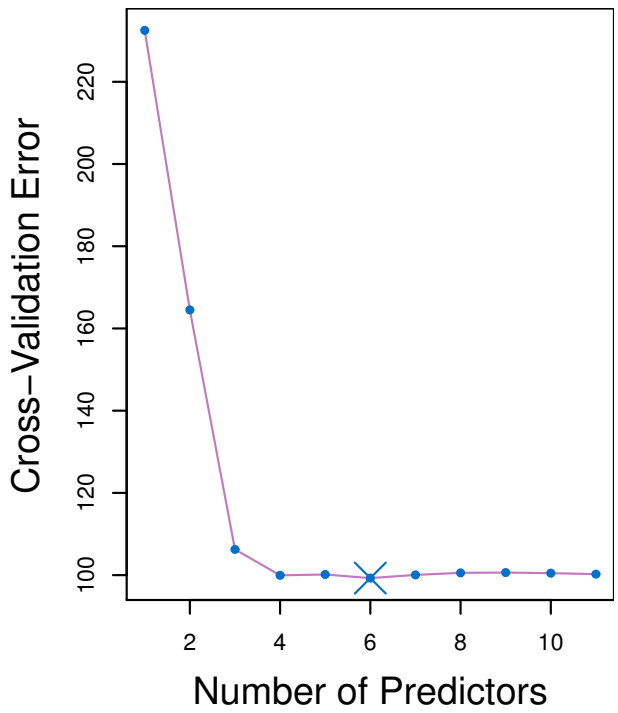
$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2)$$

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Validation and Cross-Validation (k=10)







library(ISLR); library(stargazer); stargazer(Hitters)

Statistic	N	Mean	St. Dev.	Min	Max
AtBat	263	403.643	147.307	19	687
Hits	263	107.829	45.125	1	238
HmRun	263	11.620	8.757	0	40
Runs	263	54.745	25.540	0	130
RBI	263	51.487	25.883	0	121
Walks	263	41.114	21.718	0	105
Years	263	7.312	4.794	1	24
CAtBat	263	2,657.544	2,286.583	19	14,053
CHits	263	722.186	648.200	4	4,256
CHmRun	263	69.240	82.198	0	548
CRuns	263	361.221	331.199	2	2,165
CRBI	263	330.418	323.368	3	1,659
CWalks	263	260.266	264.056	1	1,566
PutOuts	263	290.711	279.935	0	1,377
Assists	263	118.760	145.081	0	492
Errors	263	8.593	6.607	0	32
Salary	263	535.926	451.119	67.500	2,460.000

Missing Observations

```
dim(Hitters)                    322    20
sum(is.na(Hitters$Salary))       59
Hitters=na.omit(Hitters)
dim(Hitters)                    263    20
sum(is.na(Hitters))               0
```

library(leaps); regfit.full=regsubsets(Salary~.,Hitters)

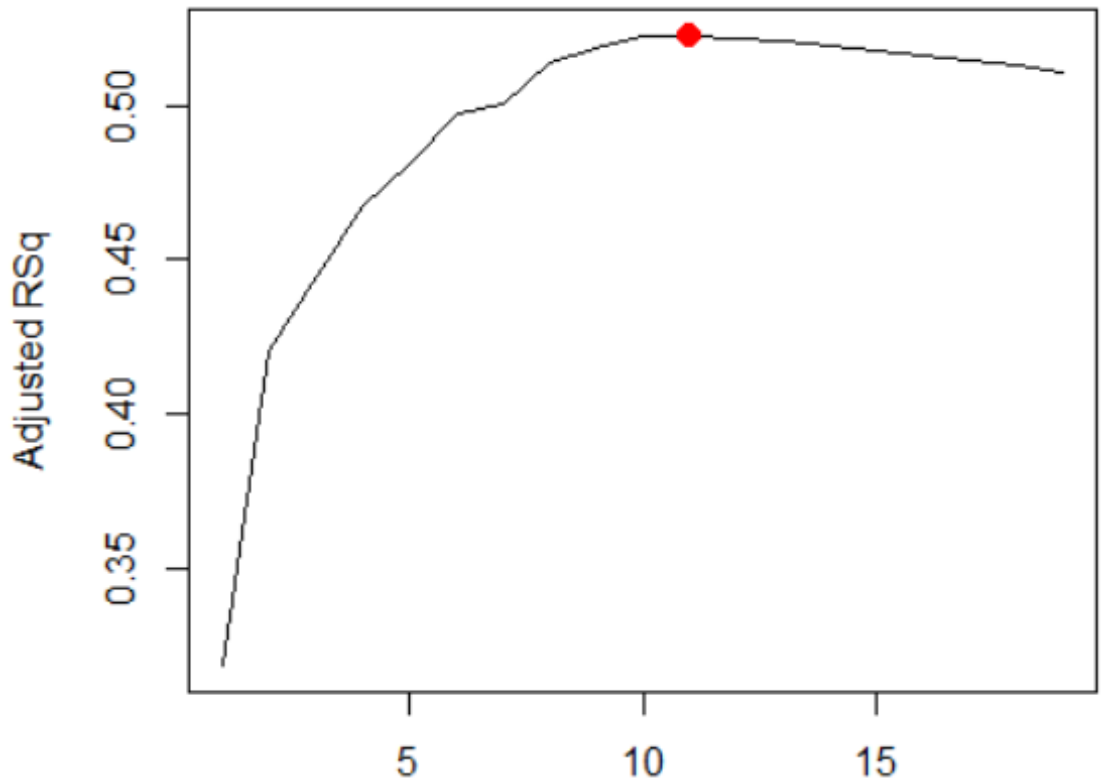
summary(regfit.full)

```
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
4  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
5  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "
7  ( 1 )  " "   "*"  " "   " "  " " "*"   " "   "*"    "*"   "*"    " "
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   "*"    "*"
          CRBI CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
1  ( 1 )  "*"  " "    " "     " "       " "     " "     " "    " "
2  ( 1 )  "*"  " "    " "     " "       " "     " "     " "    " "
3  ( 1 )  "*"  " "    " "     " "       "*"     " "     " "    " "
4  ( 1 )  "*"  " "    " "     "*"       "*"     " "     " "    " "
5  ( 1 )  "*"  " "    " "     "*"       "*"     " "     " "    " "
6  ( 1 )  "*"  " "    " "     "*"       "*"     " "     " "    " "
7  ( 1 )  " "  " "    " "     "*"       "*"     " "     " "    " "
8  ( 1 )  " "  "*"    " "     "*"       "*"     " "     " "    " "
```

plot(reg.summary$adjr2,xlab="Number of Variables", ylab="Adjusted RSq",type="l")

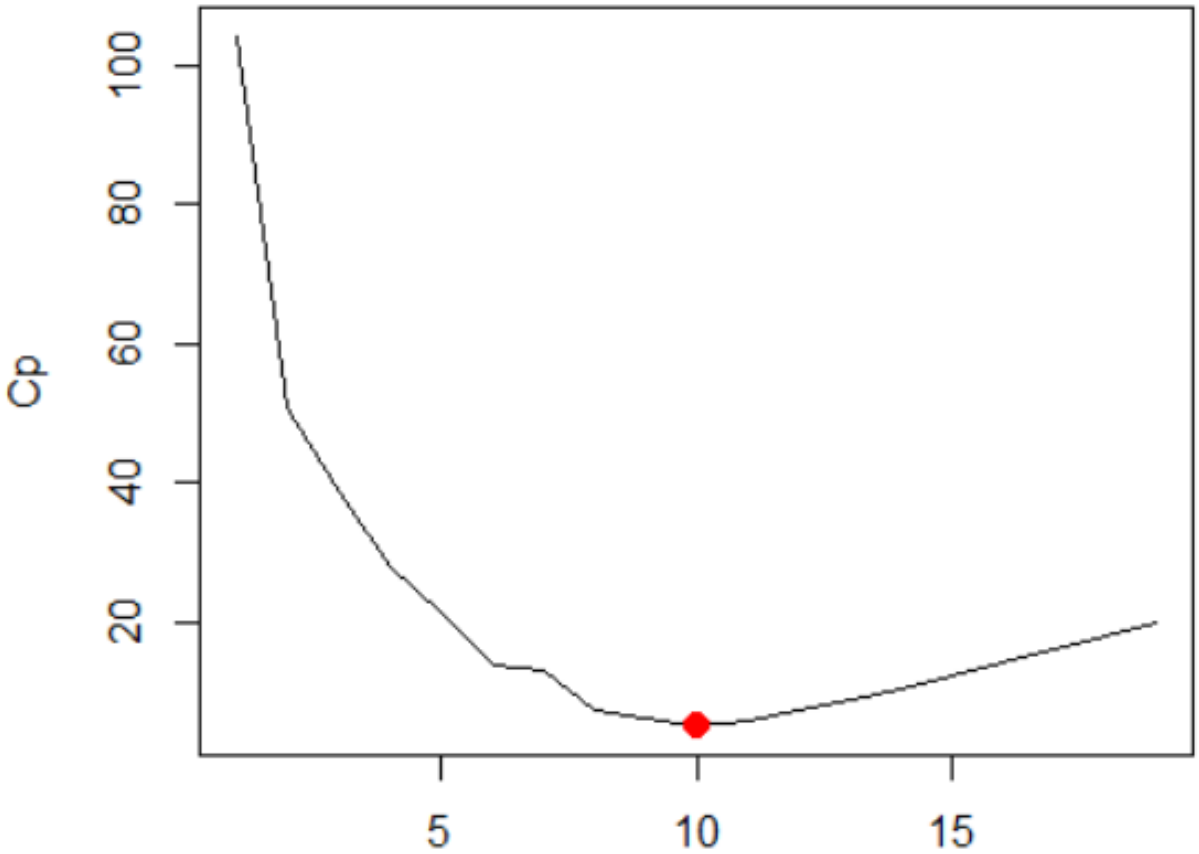
which.max(reg.summary

$adjr2)points(11,reg.summary$adjr2[11],
col="red",cex=2,pch=20)



plot(reg.summary$cp,xlab="Number of Variables", ylab="Cp",type='l')

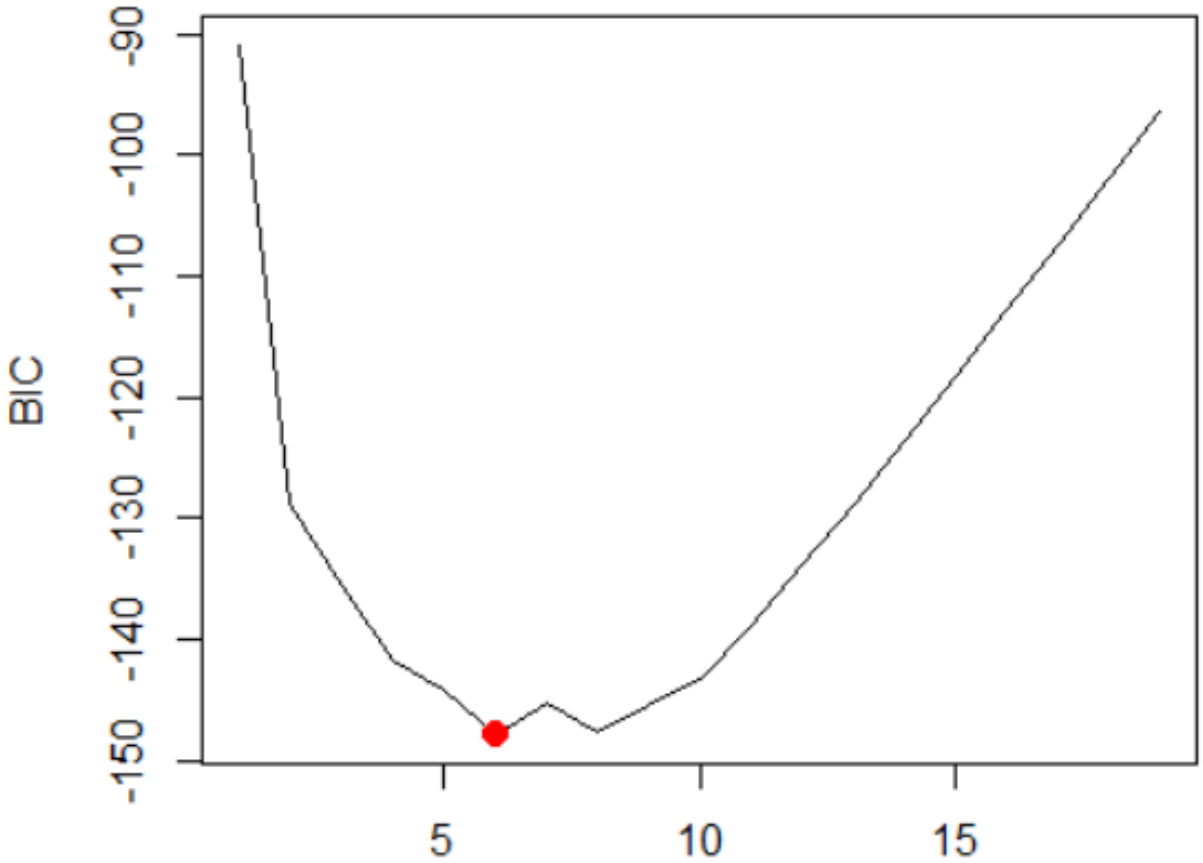
```
which.min(reg.summary$cp)
points(10,reg.summary$cp[10],col="red",cex=2,pch=20)
```



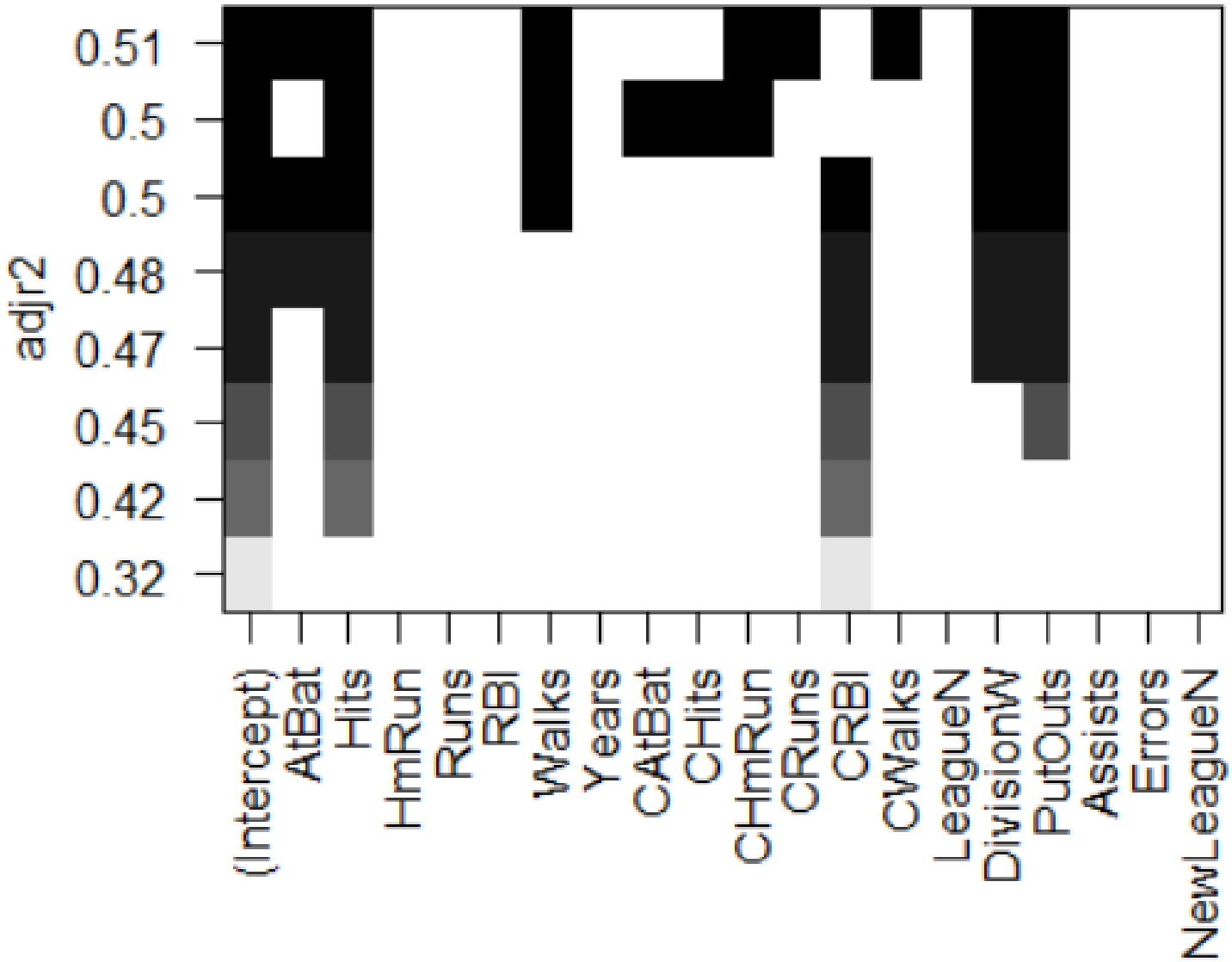
plot(reg.summary$bic,xlab="Number of Variables", ylab="BIC",type='l')

which.min(reg.summary$bic)

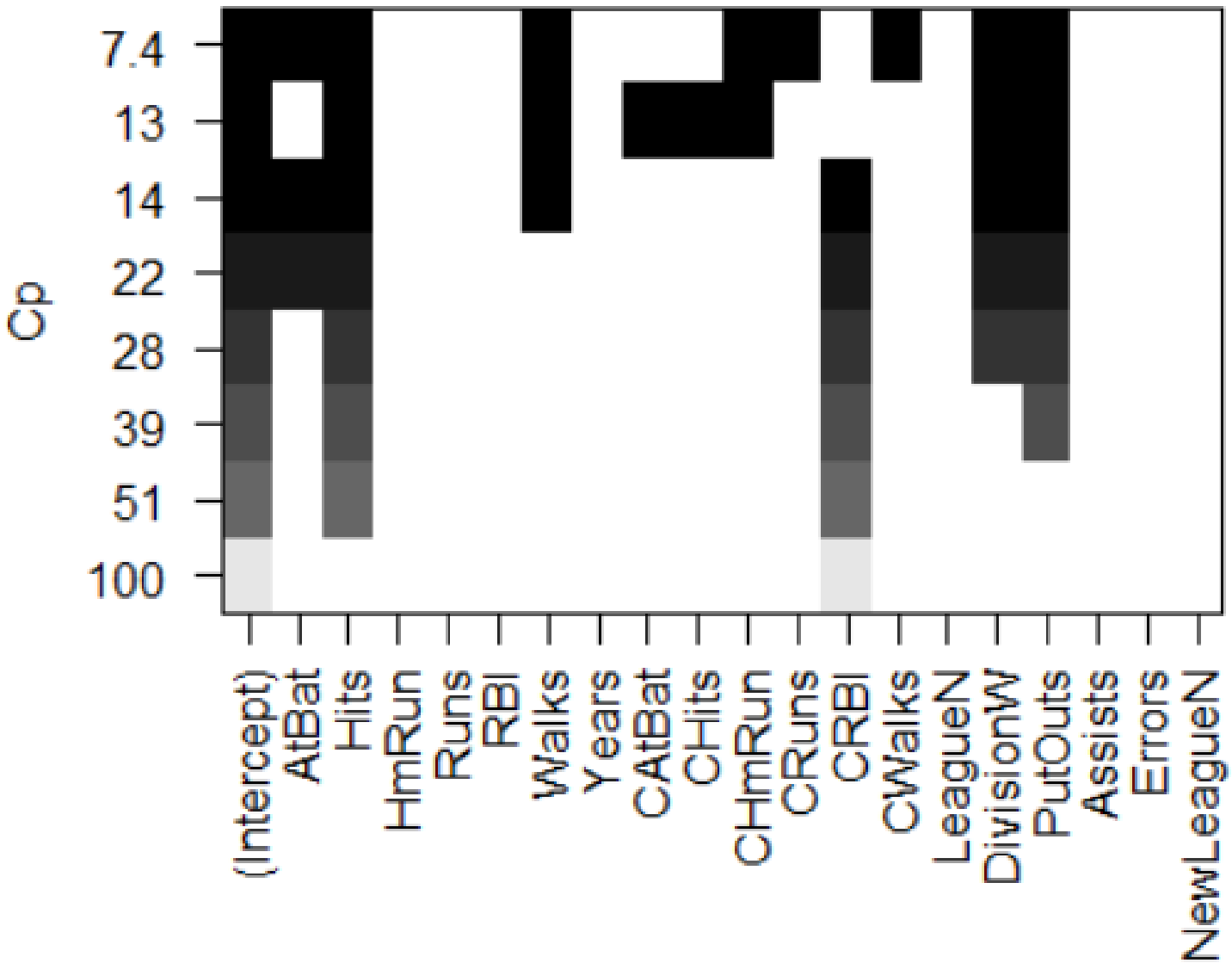
points(6,reg.summary$bic[6],col="red",cex=2,pch=20)



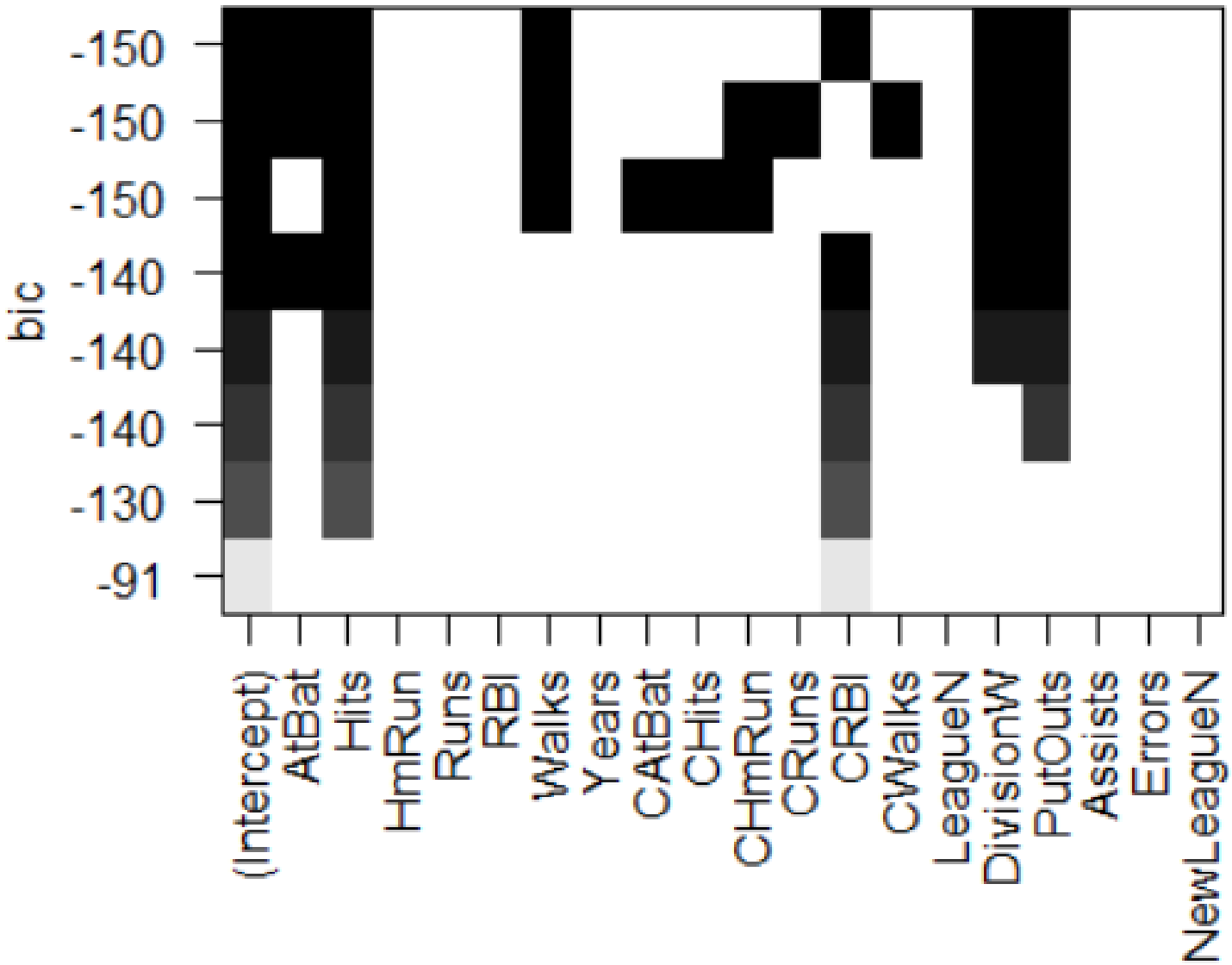
plot(regfit.full,scale="adjr2")



plot(regfit.full,scale="Cp")



plot(regfit.full,scale="bic")



regfit.fwd=regsubsets(Salary~.,data=Hitters, nvmax=9, method="forward");

summary(regfit.fwd)

```
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "
2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
4  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
5  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "
7  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "
8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"
9  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
          CRBI 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 ) "*"  "*"    " "     "*"       "*"     " "     " "    " "
```

regfit.bwd=regsubsets(Salary~.,data=Hitters, nvmax=9, method="backward")

summary(regfit.bwd)

```
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "   " "   " "   " "    "*"
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "   " "   " "   " "    "*"
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "   " "   " "   " "    "*"
4  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "   " "   " "   " "    "*"
5  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "   " "   " "   " "    "*"
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "   " "   " "   " "    "*"
7  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "   " "   " "   " "    "*"
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "   " "   " "   " "    "*"
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"   " "   " "   " "    "*"
          CRBI 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 )  "*"  "*"    " "     "*"       "*"     " "     " "    " "
```