Lab #14 - Subset Selection Methods

Econ 224

October 25th, 2018

Introduction

In this lab you will work through Section 6.5 of ISL and record your code and results in an RMarkdown document. I have added section headings below to help you organize your results. You do not have to submit this lab, so you don't have to type up a detailed description of what you've done. However, I'd suggest that you write down some notes for your own future reference. These will be helpful on the problem set. You do not need to follow the code in ISL exactly: feel free to use your preferred coding style.

You will need the `ISLR` package for the lab, so please install it if you have not done so already. This lab uses the `Hitters` dataset: in particular, we will try to predict a baseball player's `Salary` in a given year using performance statistics from the preceding year. Make sure to read the documentation file for `Hitters` before proceeding. You will also need to use the function `regsubsets` from the `leaps` library. Be sure to install `leaps` and read the documentation for `regsubsets` before proceeding. This lab also uses some "base" R plotting commands, i.e. commands that are not part of `ggplot2`. If you prefer to use `ggplot2` you are welcome to do so, but for the particular plots being made here it may be easier to stick with base R. The command `par(mfrow = c(2, 2))` is used to set up a 2×2 grid of plots for use with base R plotting functions. You can learn more about graphical parameters by entering `?par` at the R console.

Best Subset Selection

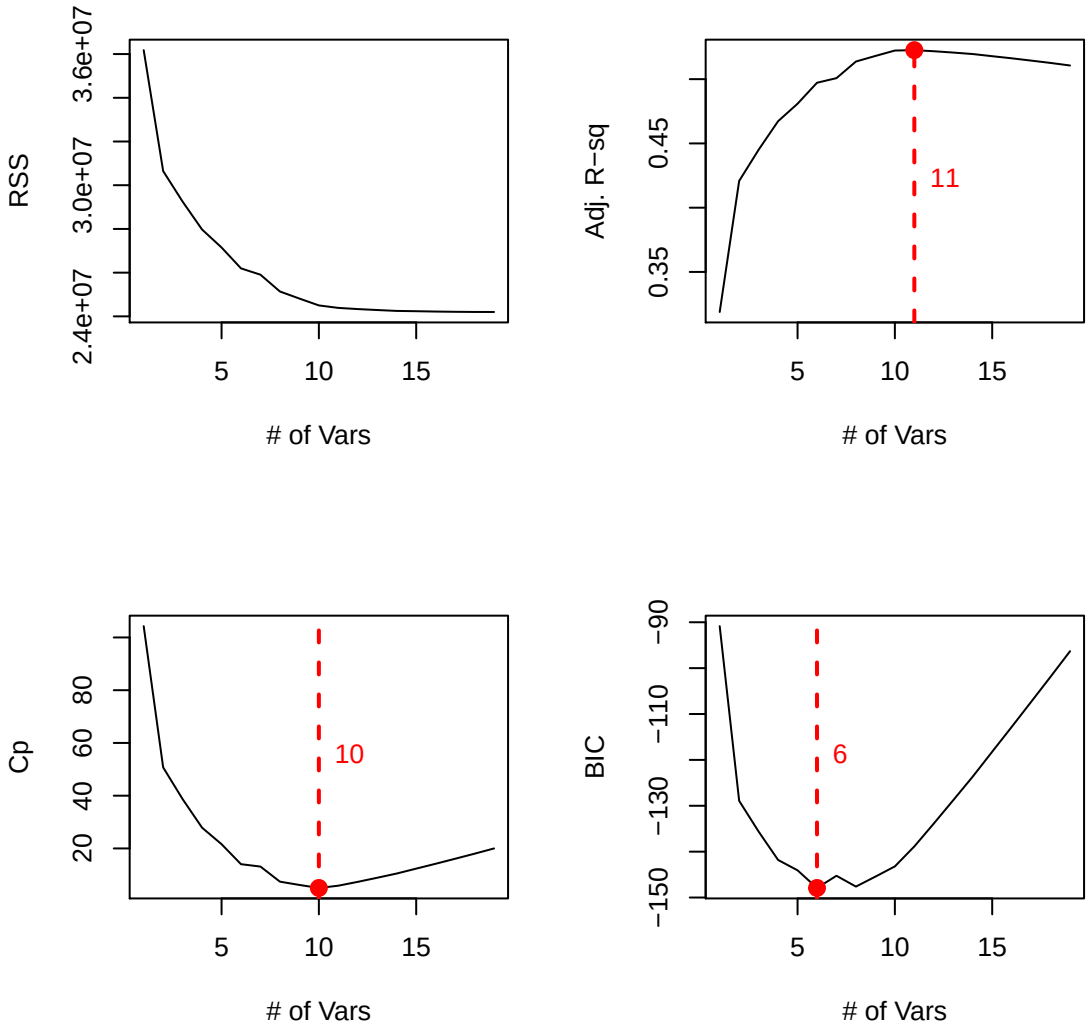
Work through section 6.5.1 of ISL and add your code and results below.

```
library(ISLR)
library(leaps)
#----------------------- Drop observations with missing data
Hitters <- na.omit(Hitters)
#----------------------- Best subsets using RSS (max of 8 vars)
reg_fit_full <- regsubsets(Salary ~ ., Hitters)
summary(reg_fit_full)
```

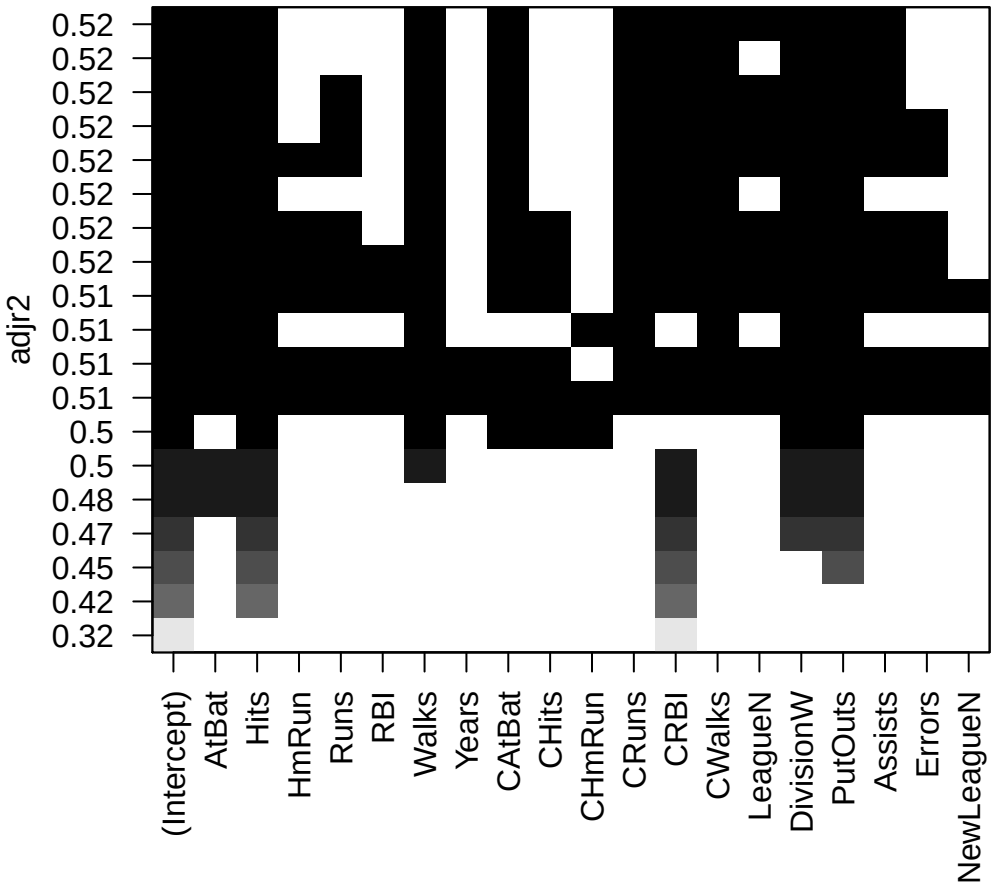
```
Subset selection object
Call: regsubsets.formula(Salary ~ ., Hitters)
19 Variables  (and intercept)
           Forced in Forced out
AtBat          FALSE      FALSE
Hits           FALSE      FALSE
HmRun          FALSE      FALSE
Runs           FALSE      FALSE
RBI            FALSE      FALSE
Walks          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
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 ) " "  "*"    " "     "*"       "*"     " "     " "    " "
```

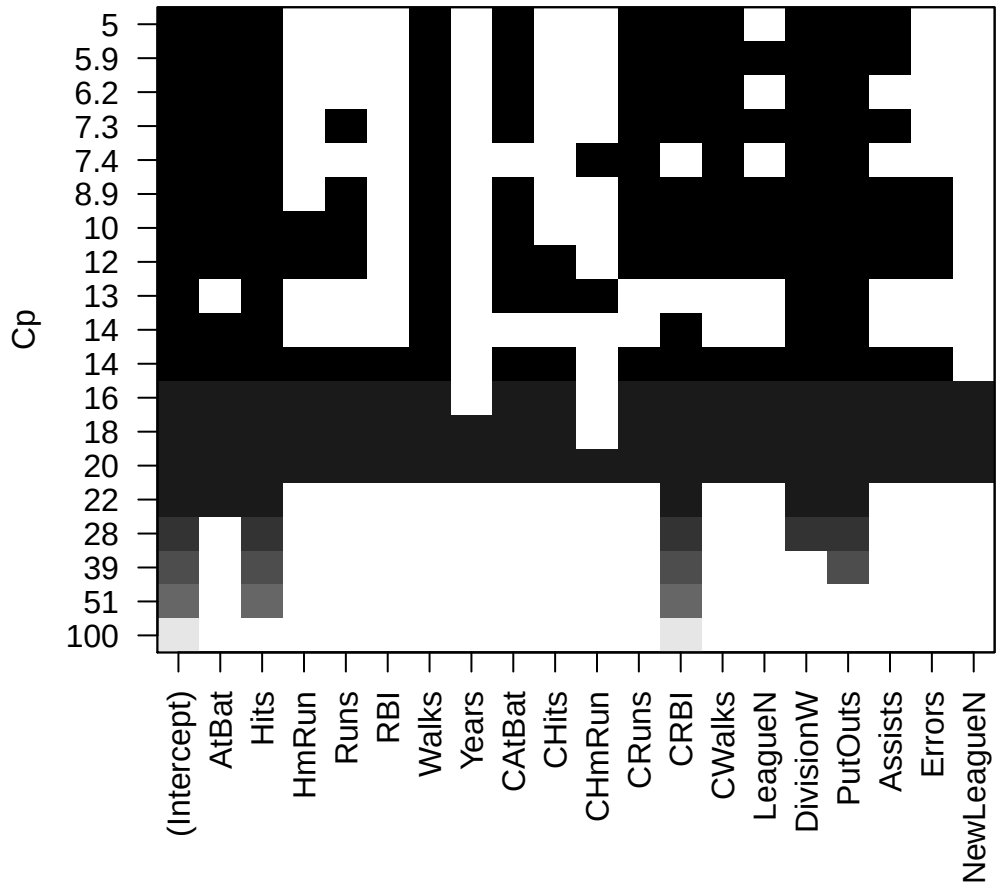
```r
#----------------------- Best subsets using RSS (no max # of vars)
reg_fit_full2 <- regsubsets(Salary ~ ., data = Hitters, nvmax = 19)
reg_summary <- summary(reg_fit_full2)
#----------------------- Function to plot min (or max) in red
plot_best <- function(subsets_summary, stat, min = TRUE) {
  y <- subsets_summary[[stat]]
  if(min) {
    x_plot <- which.min(y)
  } else {
    x_plot <- which.max(y)
  }
  abline(v = x_plot, lty = 2, col = 'red', lwd = 2)
  points(x_plot, y[x_plot], col = 'red', cex = 2, pch = 20)
  text(x_plot, median(range(y)), pos = 4, labels = x_plot, col = 'red')
}
#----------------------- Plot RSS and Adjusted R-squared
par(mfrow = c(2,2))
plot(reg_summary$rss, xlab = '# of Vars', ylab = 'RSS', type = 'l')
plot(reg_summary$adjr2, xlab = '# of Vars', ylab = 'Adj. R-sq', type = 'l')
plot_best(reg_summary, 'adjr2', FALSE)
plot(reg_summary$cp, xlab = '# of Vars', ylab = 'Cp', type = 'l')
plot_best(reg_summary, 'cp')
plot(reg_summary$bic, xlab = '# of Vars', ylab = 'BIC', type = 'l')
plot_best(reg_summary, 'bic')
```



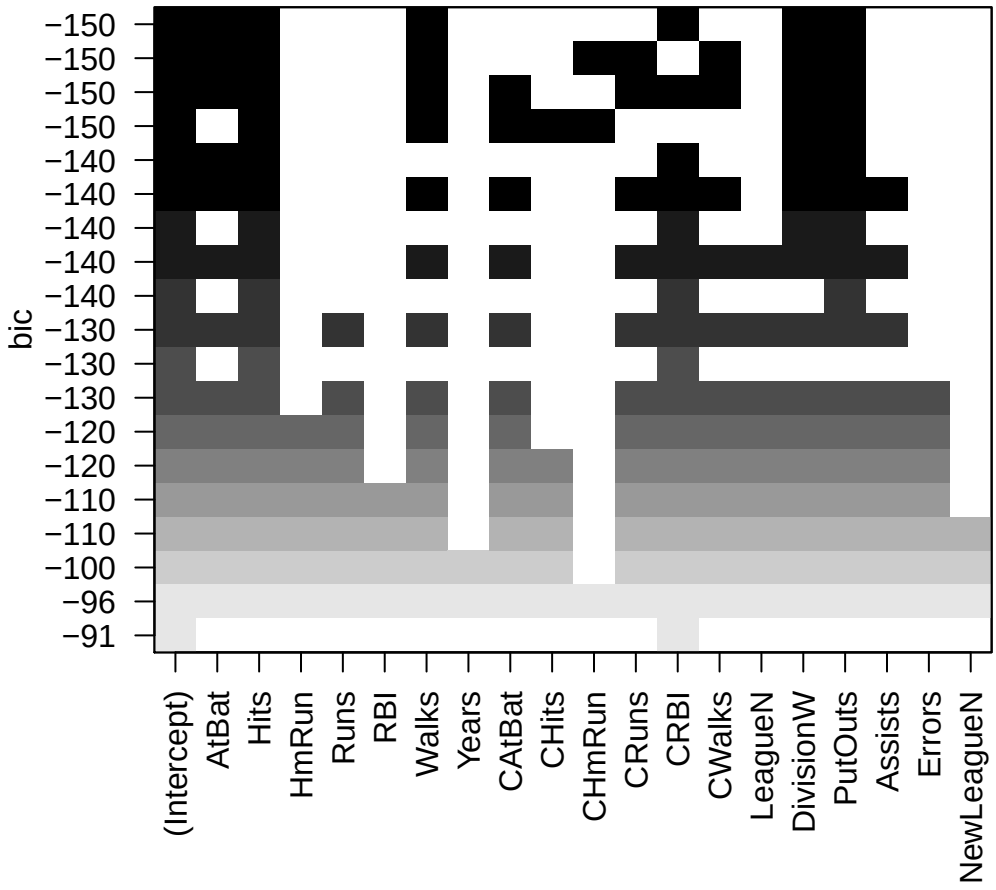

```
#---------------------- Plot best model with each number of variables (best at top)
par(mfrow = c(1, 1))
plot(reg_fit_full2, scale = 'adjr2')
```




```
plot(reg_fit_full2, scale = 'Cp')
```




```
plot(reg_fit_full2, scale = 'bic')
```




```
#----------------------- Coefficients for 6-var model with lowest RSS
coef(reg_fit_full2, 6)
```

```
 (Intercept)        AtBat         Hits        Walks         CRBI
  91.5117981   -1.8685892    7.6043976    3.6976468    0.6430169
   DivisionW      PutOuts
-122.9515338    0.2643076
```

Forward and Backward Stepwise Selection

Work through section 6.5.2 of ISL and add your code and results below.

```
#----------------------- Forward stepwise selection
reg_fit_fwd <- regsubsets(Salary ~ ., data = Hitters, nvmax = 19)
summary(reg_fit_fwd)
```

```
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 19)
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: exhaustive
          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 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"
13  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"
14  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    " "   " "    "*"
15  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    "*"   " "    "*"
16  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"
17  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"
18  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   " "    "*"
19  ( 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 )  "*"  "*"    " "     "*"       "*"     " "     " "    " "
10  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     " "    " "
11  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     " "    " "
12  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     " "    " "
13  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
14  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
15  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
```

```
16  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
17  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
18  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
19  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
```

```
#---------------------- Backward stepwise selection
reg_fit_bwd <- regsubsets(Salary ~ ., data = Hitters, nvmax = 19)
summary(reg_fit_bwd)
```

```
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 19)
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: exhaustive
          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 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"
13  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"
14  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    " "   " "    "*"
15  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    "*"   " "    "*"
16  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"
17  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"
18  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   " "    "*"
19  ( 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 )  "*"  "*"    " "     "*"       "*"     " "     " "    " "
10  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     " "    " "
11  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     " "    " "
12  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     " "    " "
13  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
14  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
15  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
16  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    " "
17  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
18  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
19  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
```

```
#---------------------- The best 7-var models differ
coef(reg_fit_fwd, 7)
```

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

```
coef(reg_fit_bwd, 7)
```

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

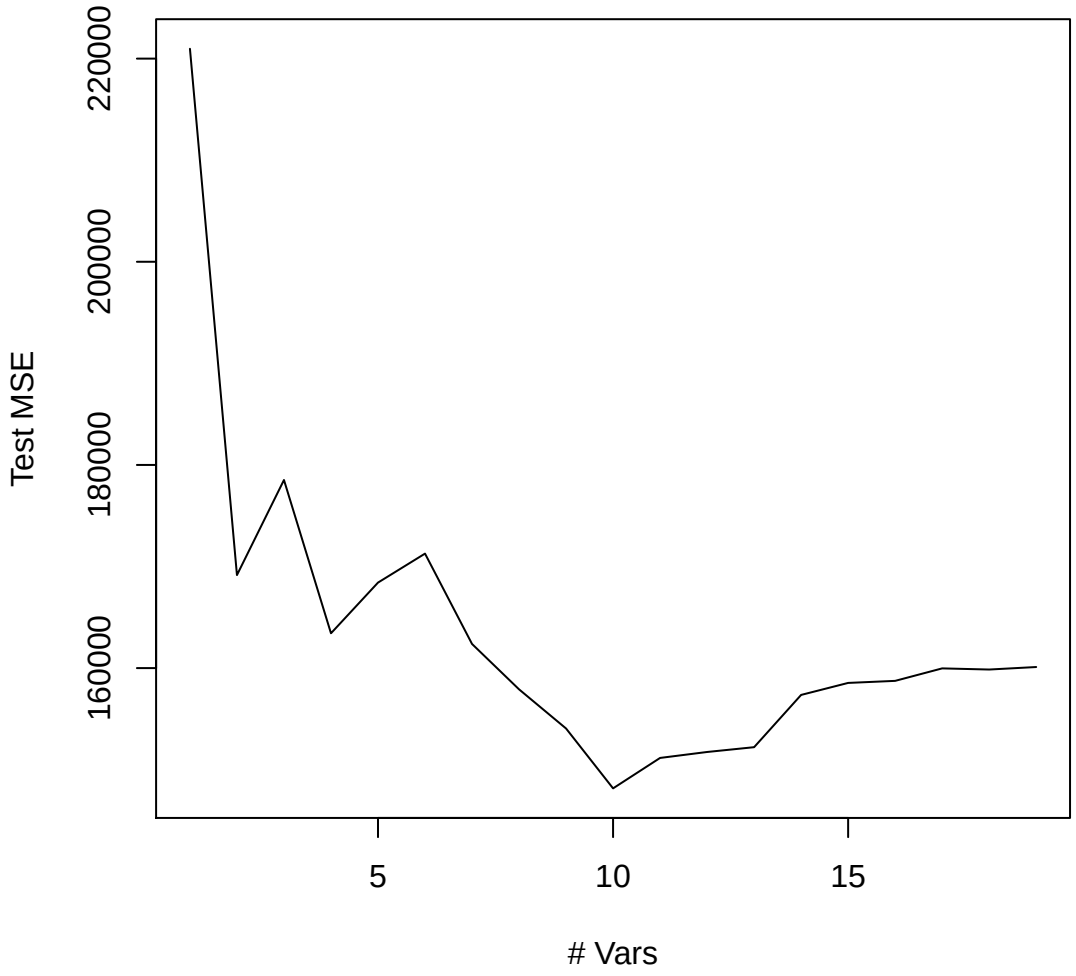
Choosing Among Models using the Validation Set Approach and CV

Work through section 6.5.3 of ISL and add your code and results below. Try to finish up through the bottom of page 252. The material on page 253 is fairly complicated and is optional.

```
#---------------------- Draw indicators for training and test subsets
set.seed(1)
train <- sample(c(TRUE, FALSE), nrow(Hitters), replace = TRUE)
test <- !train
#---------------------- Best subsets using training data
reg_fit_best <- regsubsets(Salary ~ ., data = Hitters[train,], nvmax = 19)
#---------------------- Write our own predict method for regsubsets
predict.regsubsets <- function(object, newdata, id, ...) {
  form <- as.formula(object$call[[2]])
  mat <- model.matrix(form, newdata)
  coefs <- coef(object, id = id)
```

```
  xvars <- names(coefs)
  mat[,xvars] %*% coefs
}
#---------------------- Compute test (aka validation) error for best model of each size
get_test_error <- function(i) {
  pred_i <- predict(reg_fit_best, Hitters[test,], i)
  mean((Hitters$Salary[test] - pred_i)^2)
}

#---------------------- The best model contains 10 variables
test_errors <- sapply(1:19, get_test_error)
plot(1:19, test_errors, type = 'l', xlab = '# Vars', ylab = 'Test MSE')
```




```
coef(reg_fit_best, 10)
```

```
(Intercept)       AtBat        Hits       Walks      CAtBat       CHits
-80.2751499  -1.4683816   7.1625314   3.6430345  -0.1855698   1.1053238
     CHmRun      CWalks     LeagueN   DivisionW     PutOuts
  1.3844863  -0.7483170  84.5576103 -53.0289658   0.2381662
```