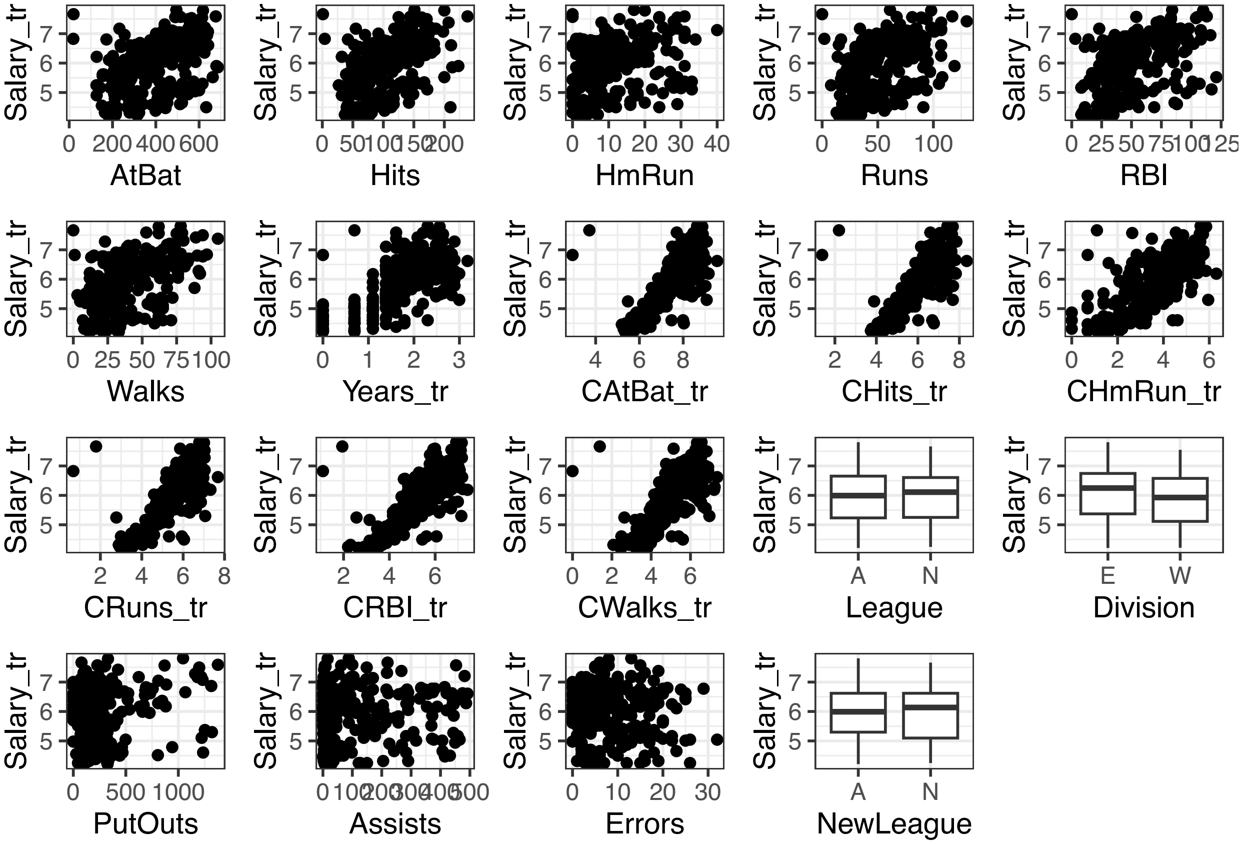
Stepwise Selection - R Code

Stepwise Selection

Example Dataset `Hitters` in ISLR2 package. Major League Baseball Data from the 1986 and 1987 seasons.

References James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, https://www.statlearning.com, Springer-Verlag, New York

Load and look at the data



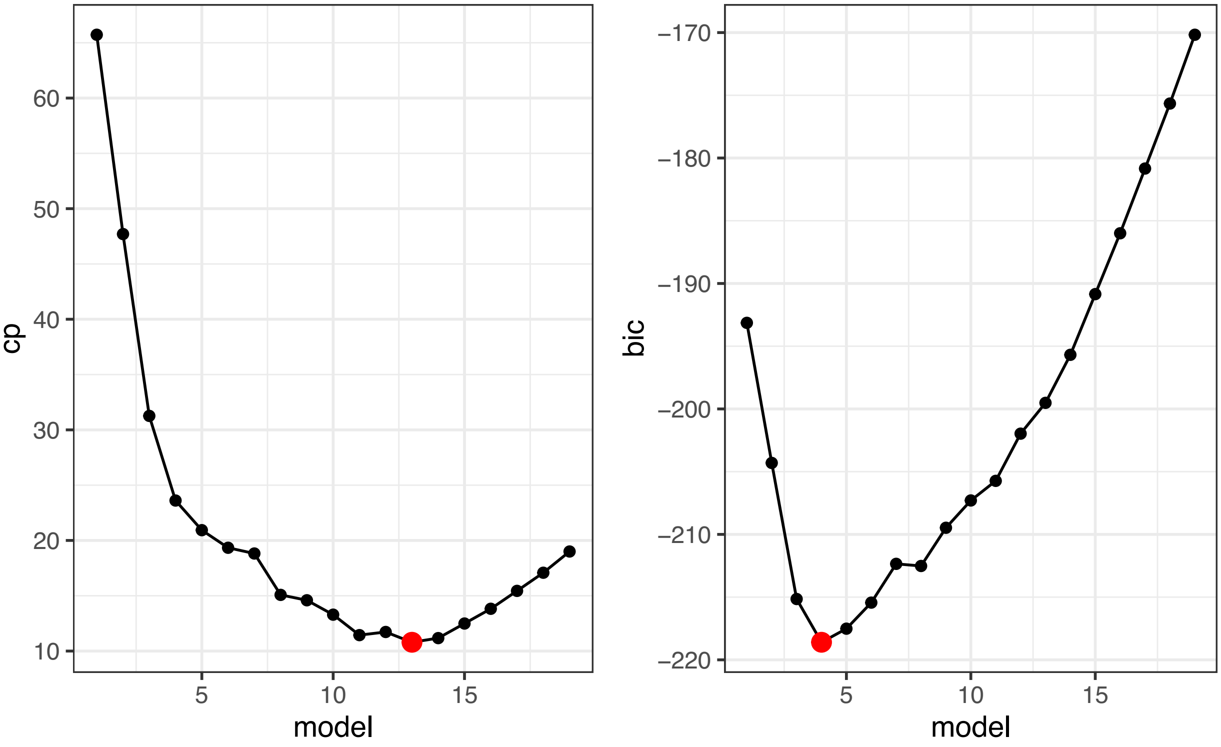
Forward stepwise selection

Here we use the `regsubsets` function but specify the `method="forward"` option.

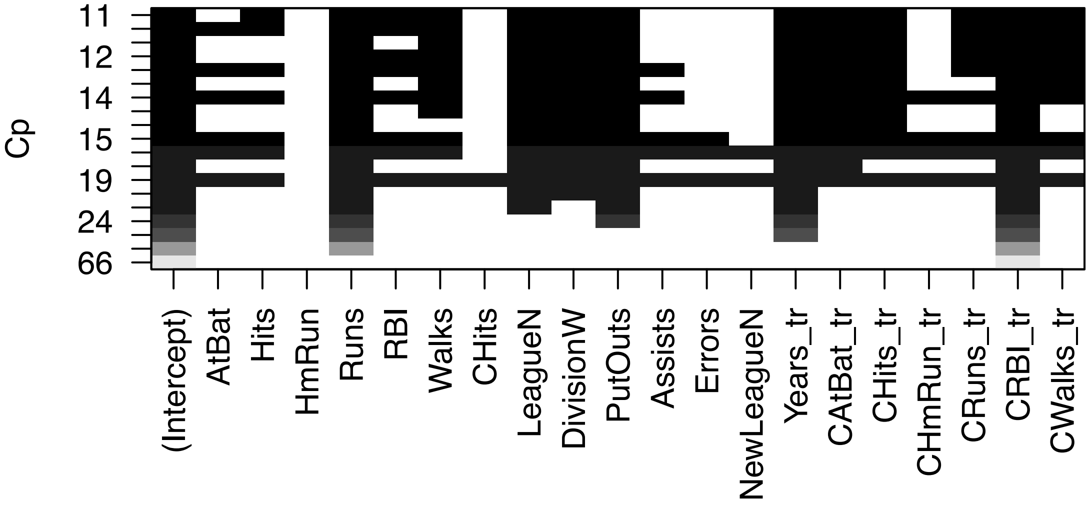
```
regfit.fwd <- regsubsets(Salary_tr ~ ., Hitters_tr, nvmax = 19, method = "forward")
# should specify the method
```

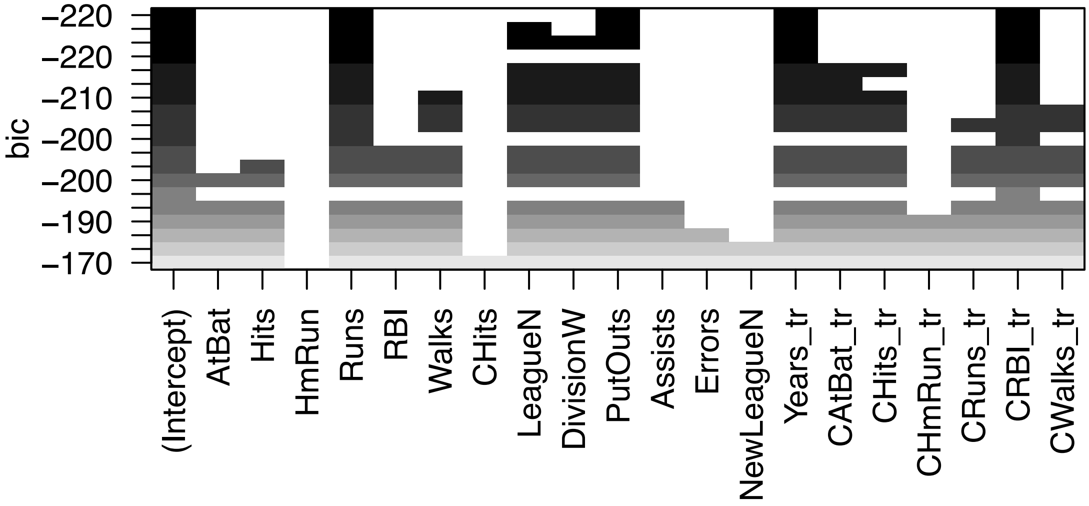
```
results.fwd <- data.frame(model = seq(1,19),
                          cp = summary(regfit.fwd)$cp,
                          bic = summary(regfit.fwd)$bic)
p_cp <- ggplot(results.fwd, aes(x = model, y = cp)) +
  geom_point() + geom_line() + theme_bw() +
  geom_point(data = results.fwd[which.min(summary(regfit.fwd)$cp),],
              aes(y = cp, x = model),
              size = 3, col = "red")
p_bic <- ggplot(results.fwd, aes(x = model, y = bic)) +
  geom_point() + geom_line() + theme_bw() +
  geom_point(data = results.fwd[which.min(summary(regfit.fwd)$bic),],
              aes(y = bic, x = model),
              size = 3, col = "red")
grid.arrange(p_cp, p_bic, ncol =2)
```

Compare model selection with forward stepwise based on Mallow's Cp and BIC



```r
par(mfrow=c(2,1))
plot(regfit.fwd, scale="Cp")
plot(regfit.fwd, scale="bic")
```





Get coefficients of “Best models” Based on Cp

```r
# The best model based on Cp is the one with 13 variables
coef(regfit.fwd, 13)
```

```
##   (Intercept)          Hits          Runs           RBI         Walks 
##  6.1480388829  0.0046359986 -0.0005457976 -0.0062465530  0.0099115302 
##       LeagueN     DivisionW       PutOuts      Years_tr     CAtBat_tr 
##  0.1227861620 -0.1335796448  0.0003101371  0.6079919422 -1.2451991830 
##      CHits_tr      CRuns_tr       CRBI_tr     CWalks_tr 
##  0.4154311743  0.6833706603  0.5941897634 -0.3980564299
```

Based on BIC

```
# The best model based on BIC is the one with 4 variables
coef(regfit.fwd, 4)
```

```
##  (Intercept)         Runs      PutOuts     Years_tr      CRBI_tr
## 3.3116818532 0.0090245142 0.0003881659 0.4452743977 0.2339874627
```

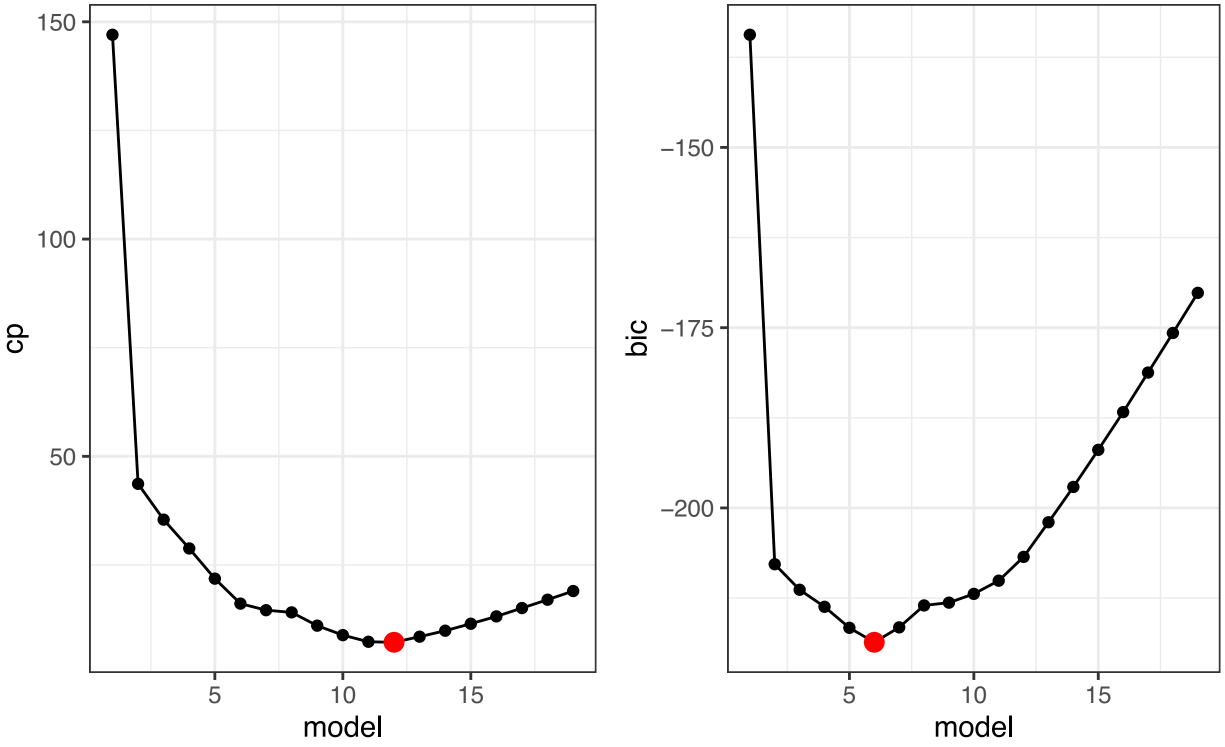
Backward stepwise selection

Repeat this exercise with the backward stepwise selection. Which model would you pick? Which variables are included in the model? Do you find the same results as with the forward stepwise selection method?

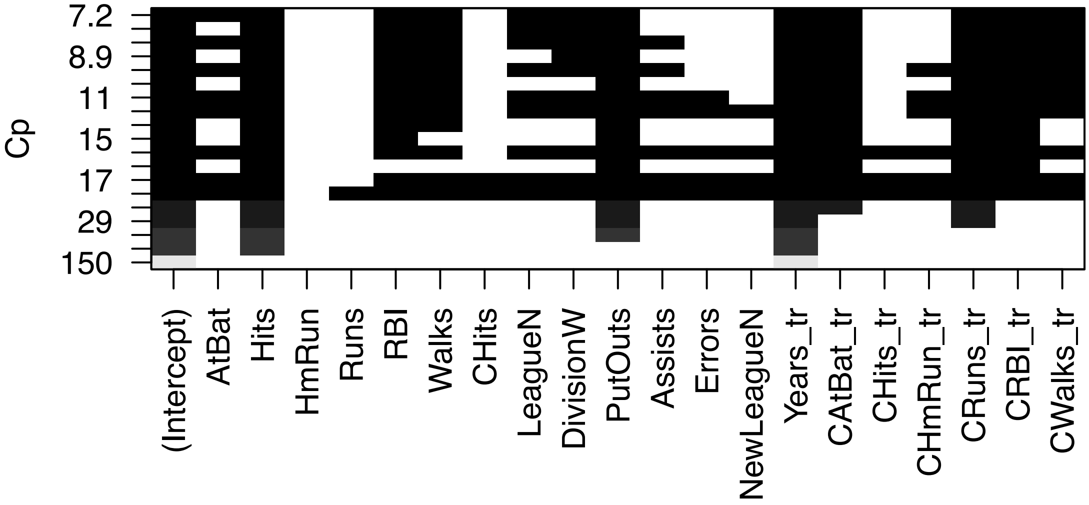
```
regfit.bwd <- regsubsets(Salary_tr ~ ., Hitters_tr, nvmax = 19, method = "backward")
```

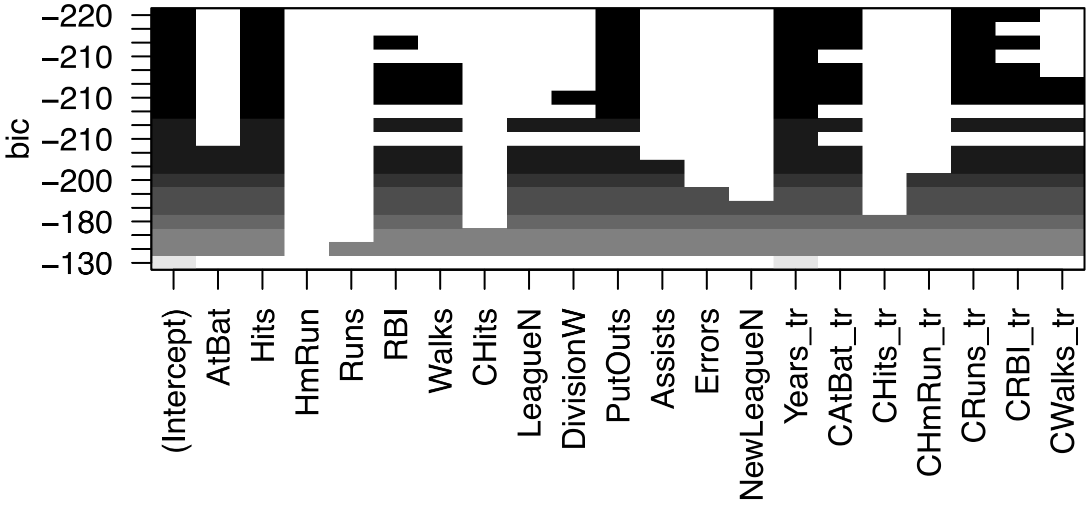
```
results.bwd <- data.frame(model = seq(1,19),
                          cp = summary(regfit.bwd)$cp,
                          bic = summary(regfit.bwd)$bic)
p_cp <- ggplot(results.bwd, aes(x = model, y = cp)) +
  geom_point() + geom_line() + theme_bw() +
  geom_point(data = results.bwd[which.min(summary(regfit.bwd)$cp),],
             aes(y = cp, x = model),
             size = 3, col = "red")
p_bic <- ggplot(results.bwd, aes(x = model, y = bic)) +
  geom_point() + geom_line() + theme_bw() +
  geom_point(data = results.bwd[which.min(summary(regfit.bwd)$bic),],
             aes(y = bic, x = model),
             size = 3, col = "red")
grid.arrange(p_cp, p_bic, ncol =2)
```

Compare model selection with backward stepwise based on Mallow's Cp and BIC



```
par(mfrow=c(2,1))
plot(regfit.bwd, scale="Cp")
plot(regfit.bwd, scale="bic")
```





Get coefficients of "Best models" Based on Cp

```
# The best model based on Cp is the one with 12 variables
coef(regfit.bwd, 12)
```

```
##   (Intercept)         AtBat          Hits           RBI         Walks
##  5.2265488492 -0.0015019022  0.0085462263 -0.0054814514  0.0110060856
##       LeagueN     DivisionW       PutOuts      Years_tr     CAtBat_tr
##  0.1290951738 -0.1283650435  0.0003237136  0.5619165175 -0.7179913661
##      CRuns_tr       CRBI_tr     CWalks_tr
##  0.6915021633  0.5693710251 -0.4458114863
```

Based on BIC

```
# The best model based on BIC is the one with 6 variables
coef(regfit.bwd, 6)
```

```
##   (Intercept)          Hits       PutOuts      Years_tr     CAtBat_tr 
##  5.7588083701  0.0049922574  0.0003675681  0.5945822178 -0.9345365295 
##      CRuns_tr       CRBI_tr 
##  0.6804049770  0.3361272177
```

Do you find the same results as with the forward stepwise selection method?

The models selected with the Cp criterion are very similar. The forward stepwise selection has 13 variables and the backward, 12. However, the forward stepwise selection method includes the additional variables `Runs` and `CHits_tr` and does not include the variable `AtBat`.

The models selected with the BIC criterion are also fairly similar. They have a similar number of variables (6 and 4), and have three variables in common, i.e. `PutOuts`, `Years_tr` and `CRBI_tr`.