2022년 2학기

Computational Statistics

HW#3

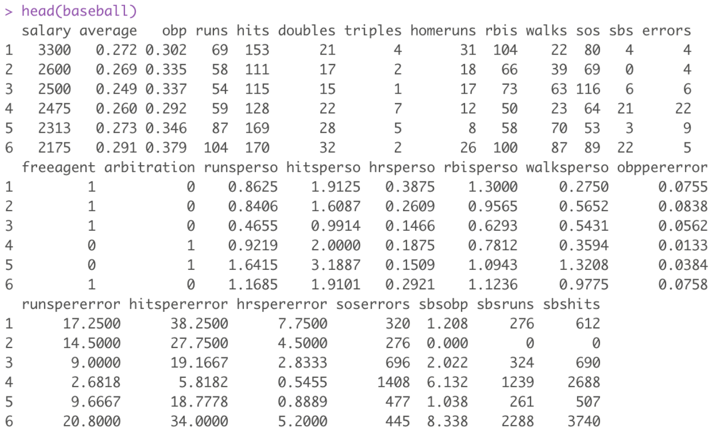
222STG10

김희숙

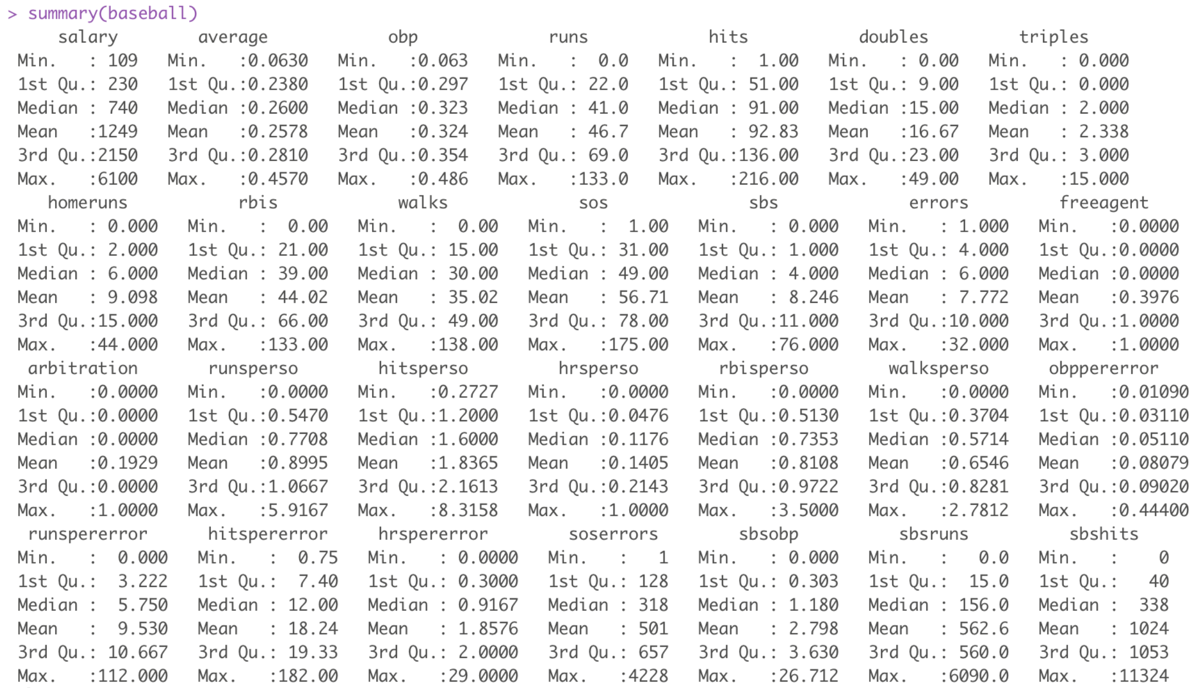
Baseball salary data

1. All possible regression + AIC

baseball.dat.txt는 337 x 28 인 data로 head()와 summary()를 통해 data를 보면 다음과 같다.

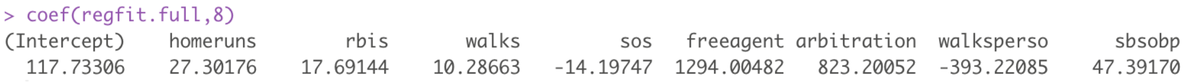


```
> head(baseball)
  salary average   obp runs hits doubles triples homeruns rbis walks sos sbs errors
1   3300   0.272 0.302   69  153      21       4       31  104    22  80   4      4
2   2600   0.269 0.335   58  111      17       2       18   66    39  69   0      4
3   2500   0.249 0.337   54  115      15       1       17   73    63 116   6      6
4   2475   0.260 0.292   59  128      22       7       12   50    23  64  21     22
5   2313   0.273 0.346   87  169      28       5        8   58    70  53   3      9
6   2175   0.291 0.379  104  170      32       2       26  100    87  89  22      5
  freeagent arbitration runsperso hitsperso hrsperso rbisperso walksperso obppererror
1         1           0    0.8625    1.9125   0.3875    1.3000     0.2750      0.0755
2         1           0    0.8406    1.6087   0.2609    0.9565     0.5652      0.0838
3         1           0    0.4655    0.9914   0.1466    0.6293     0.5431      0.0562
4         0           1    0.9219    2.0000   0.1875    0.7812     0.3594      0.0133
5         0           1    1.6415    3.1887   0.1509    1.0943     1.3208      0.0384
6         1           0    1.1685    1.9101   0.2921    1.1236     0.9775      0.0758
  runspererror hitspererror hrspererror soserrors sbsobp sbsruns sbshits
1      17.2500      38.2500      7.7500       320  1.208     276     612
2      14.5000      27.7500      4.5000       276  0.000       0       0
3       9.0000      19.1667      2.8333       696  2.022     324     690
4       2.6818       5.8182      0.5455      1408  6.132    1239    2688
5       9.6667      18.7778      0.8889       477  1.038     261     507
6      20.8000      34.0000      5.2000       445  8.338    2288    3740
```



```
> summary(baseball)
     salary         average              obp              runs            hits            doubles          triples      
 Min.   : 109   Min.   :0.0630   Min.   :0.063   Min.   :  0.0   Min.   :  1.00   Min.   : 0.00   Min.   : 0.000  
 1st Qu.: 230   1st Qu.:0.2380   1st Qu.:0.297   1st Qu.: 22.0   1st Qu.: 51.00   1st Qu.: 9.00   1st Qu.: 0.000  
 Median : 740   Median :0.2600   Median :0.323   Median : 41.0   Median : 91.00   Median :15.00   Median : 2.000  
 Mean   :1249   Mean   :0.2578   Mean   :0.324   Mean   : 46.7   Mean   : 92.83   Mean   :16.67   Mean   : 2.338  
 3rd Qu.:2150   3rd Qu.:0.2810   3rd Qu.:0.354   3rd Qu.: 69.0   3rd Qu.:136.00   3rd Qu.:23.00   3rd Qu.: 3.000  
 Max.   :6100   Max.   :0.4570   Max.   :0.486   Max.   :133.0   Max.   :216.00   Max.   :49.00   Max.   :15.000  
    homeruns          rbis            walks             sos              sbs             errors         freeagent     
 Min.   : 0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  1.00   Min.   : 0.000   Min.   : 1.000   Min.   :0.0000  
 1st Qu.: 2.000   1st Qu.: 21.00   1st Qu.: 15.00   1st Qu.: 31.00   1st Qu.: 1.000   1st Qu.: 4.000   1st Qu.:0.0000  
 Median : 6.000   Median : 39.00   Median : 30.00   Median : 49.00   Median : 4.000   Median : 6.000   Median :0.0000  
 Mean   : 9.098   Mean   : 44.02   Mean   : 35.02   Mean   : 56.71   Mean   : 8.246   Mean   : 7.772   Mean   :0.3976  
 3rd Qu.:15.000   3rd Qu.: 66.00   3rd Qu.: 49.00   3rd Qu.: 78.00   3rd Qu.:11.000   3rd Qu.:10.000   3rd Qu.:1.0000  
 Max.   :44.000   Max.   :133.00   Max.   :138.00   Max.   :175.00   Max.   :76.000   Max.   :32.000   Max.   :1.0000  
  arbitration       runsperso        hitsperso         hrsperso        rbisperso        walksperso      obppererror     
 Min.   :0.0000   Min.   :0.0000   Min.   :0.2727   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.01090  
 1st Qu.:0.0000   1st Qu.:0.5470   1st Qu.:1.2000   1st Qu.:0.0476   1st Qu.:0.5130   1st Qu.:0.3704   1st Qu.:0.03110  
 Median :0.0000   Median :0.7708   Median :1.6000   Median :0.1176   Median :0.7353   Median :0.5714   Median :0.05110  
 Mean   :0.1929   Mean   :0.8995   Mean   :1.8365   Mean   :0.1405   Mean   :0.8108   Mean   :0.6546   Mean   :0.08079  
 3rd Qu.:0.0000   3rd Qu.:1.0667   3rd Qu.:2.1613   3rd Qu.:0.2143   3rd Qu.:0.9722   3rd Qu.:0.8281   3rd Qu.:0.09020  
 Max.   :1.0000   Max.   :5.9167   Max.   :8.3158   Max.   :1.0000   Max.   :3.5000   Max.   :2.7812   Max.   :0.44400  
  runspererror      hitspererror     hrspererror        soserrors          sbsobp          sbsruns          sbshits     
 Min.   :  0.000   Min.   :  0.75   Min.   : 0.0000   Min.   :   1   Min.   : 0.000   Min.   :   0.0   Min.   :    0  
 1st Qu.:  3.222   1st Qu.:  7.40   1st Qu.: 0.3000   1st Qu.: 128   1st Qu.: 0.303   1st Qu.:  15.0   1st Qu.:   40  
 Median :  5.750   Median : 12.00   Median : 0.9167   Median : 318   Median : 1.180   Median : 156.0   Median :  338  
 Mean   :  9.530   Mean   : 18.24   Mean   : 1.8576   Mean   : 501   Mean   : 2.798   Mean   : 562.6   Mean   : 1024  
 3rd Qu.: 10.667   3rd Qu.: 19.33   3rd Qu.: 2.0000   3rd Qu.: 657   3rd Qu.: 3.630   3rd Qu.: 560.0   3rd Qu.: 1053  
 Max.   :112.000   Max.   :182.00   Max.   :29.0000   Max.   :4228   Max.   :26.712   Max.   :6090.0   Max.   :11324  
```

All possible regression을 leaps library의 regsubsets함수를 통해 구해본 후, which.min를 통해 AIC가 가장 작은 모형을 알아보면 다음과 같다.



```
> coef(regfit.full,8)
(Intercept)    homeruns        rbis       walks         sos   freeagent arbitration  walksperso      sbsobp 
  117.73306    27.30176    17.69144    10.28663   -14.19747  1294.00482   823.20052  -393.22085    47.39170 
```

즉, salary = 117.73306 + 27.30176* homeruns + 17.69144* rbis + 10.28663* walks + -14.19747* sos + 1294.00482* freeagent + 823.20052* arbitration + -393.22085* walksperso + 47.39170* sbsobp

모형이 all possible regression을 AIC를 기준으로 비교하였을 때 제일 best인 model이다. 이를 lm을 이용해 모델 정의 후 summary() 및 AIC를 구해보면 다음과 같다.

```
> summary(mylm)

Call:
lm(formula = salary ~ homeruns + rbis + walks + sos + freeagent +
    arbitration + walksperso + sbsobp, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-2035.0  -460.6    41.5   357.0  2944.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.733    134.449   0.876  0.38185
homeruns      27.302      9.378   2.911  0.00385 **
rbis          17.691      3.167   5.587 4.87e-08 ***
walks         10.287      3.844   2.676  0.00782 **
sos          -14.197      2.582  -5.498 7.72e-08 ***
freeagent   1294.005     94.040  13.760  < 2e-16 ***
arbitration  823.201    110.444   7.454 8.15e-13 ***
walksperso  -393.221    173.936  -2.261  0.02443 *
sbsobp        47.392     10.399   4.557 7.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 692.5 on 328 degrees of freedom
Multiple R-squared:  0.6956,	Adjusted R-squared:  0.6882
F-statistic: 93.69 on 8 and 328 DF,  p-value: < 2.2e-16

> extractAIC(mylm)
[1]    9.000 4416.997
```

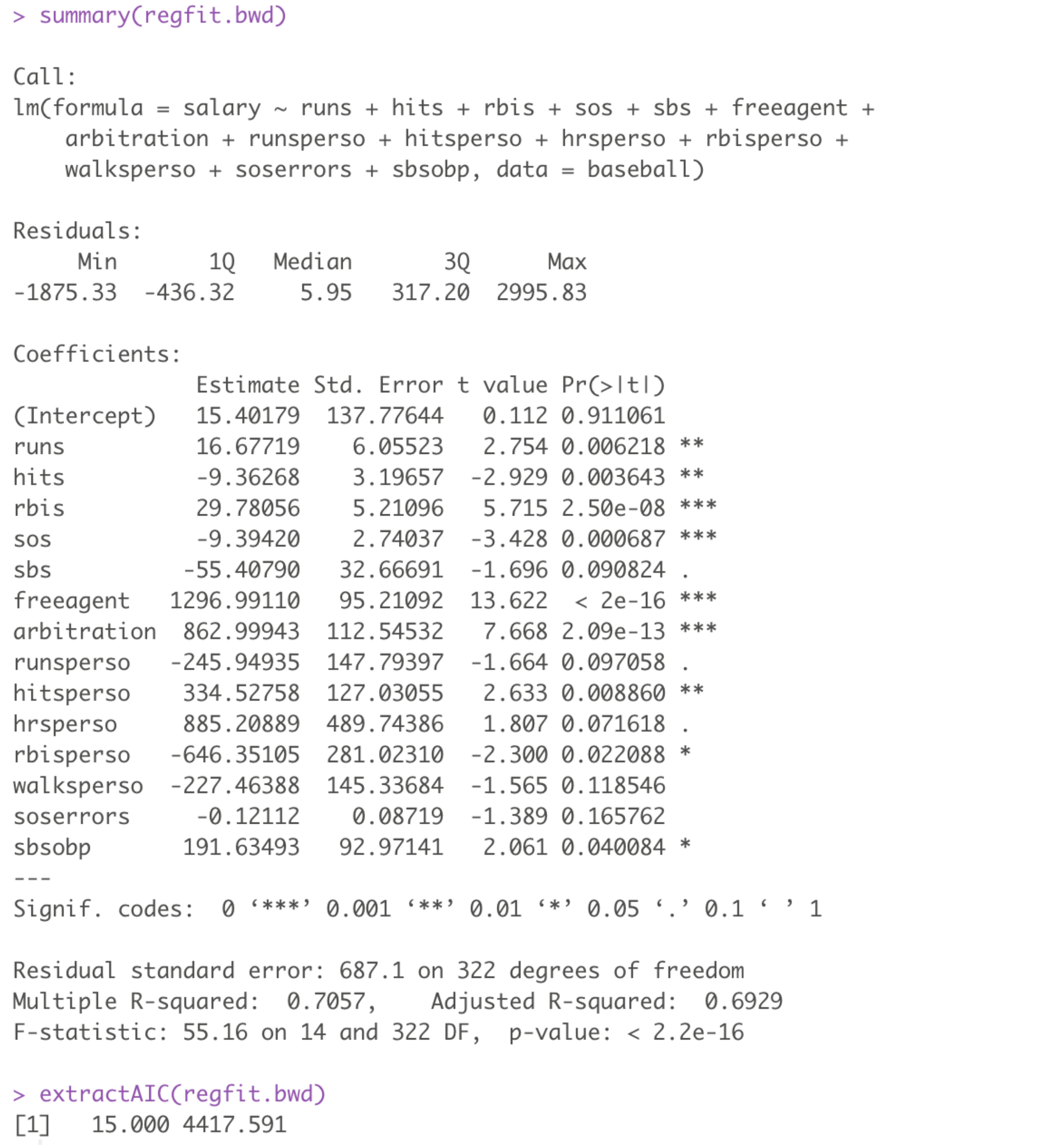
2. Stepwise

Full모형에 대해 stepwise를 Forward, Backward and Both selection을 진행한 결과는 다음과 같다.

Forward stepwise selection

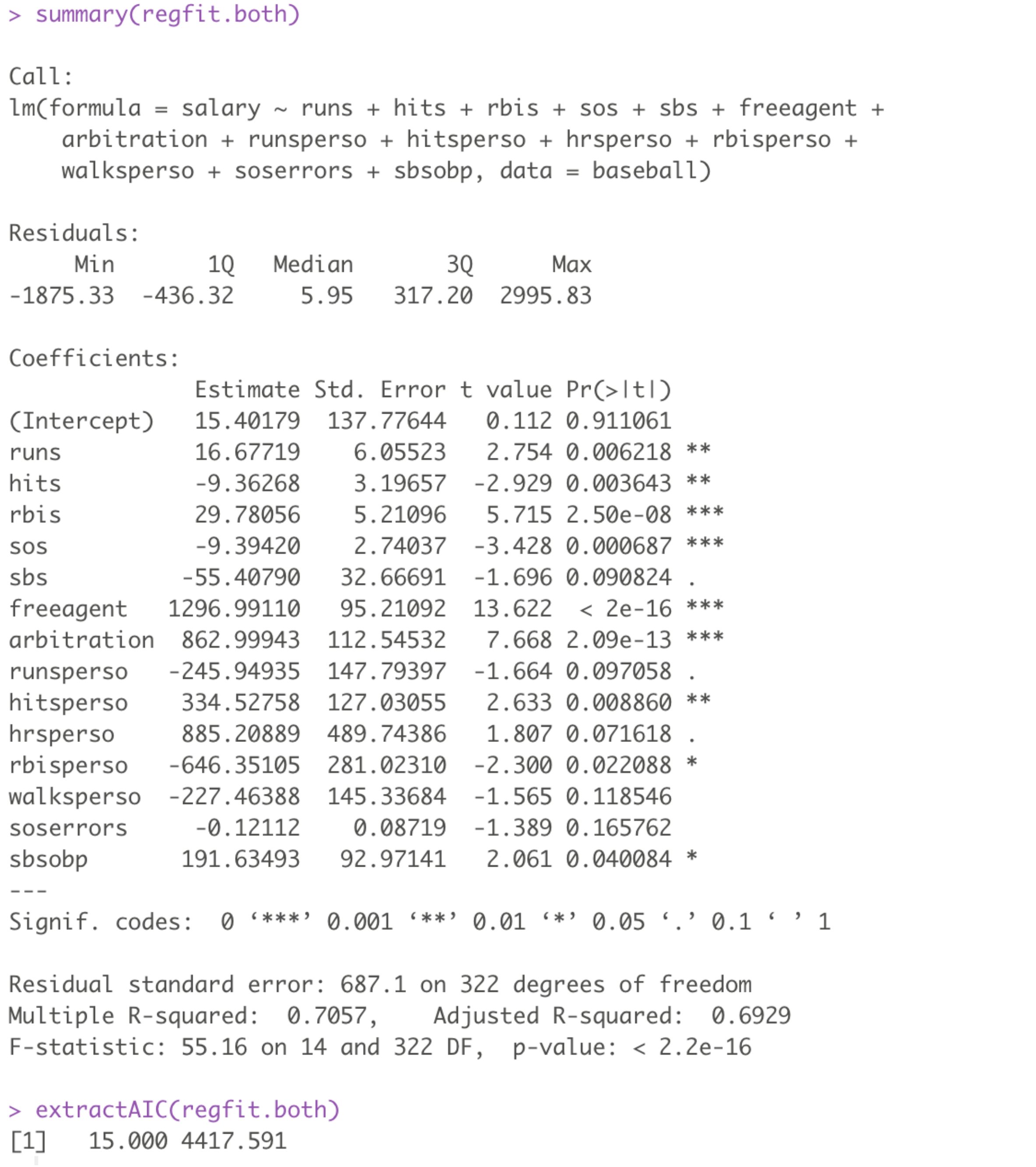
```
> summary(regfit.fwd)

Call:
lm(formula = salary ~ rbis + freeagent + arbitration + sbsruns +
    sos + homeruns + rbisperso + soserrors + runs, data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-1904.13  -443.35    26.77   324.10  3035.58

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.84531  131.14043   0.205  0.83793
rbis          19.51799    4.60477   4.239 2.93e-05 ***
freeagent   1276.97086   93.83128  13.609  < 2e-16 ***
arbitration  814.64451  111.52486   7.305 2.14e-12 ***
sbsruns        0.15612    0.05341   2.923  0.00371 **
sos          -10.80154    2.69939  -4.001 7.79e-05 ***
homeruns      24.51855    9.51986   2.576  0.01045 *
rbisperso   -208.48832  130.16519  -1.602  0.11018
soserrors     -0.13431    0.08765  -1.532  0.12640
runs           5.04389    3.29855   1.529  0.12720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 694 on 327 degrees of freedom
Multiple R-squared:  0.6951,	Adjusted R-squared:  0.6867
F-statistic: 82.84 on 9 and 327 DF,  p-value: < 2.2e-16

> extractAIC(regfit.fwd)
[1]   10.000 4419.512
```

Backward stepwise selection



```
> summary(regfit.bwd)

Call:
lm(formula = salary ~ runs + hits + rbis + sos + sbs + freeagent +
    arbitration + runsperso + hitsperso + hrsperso + rbisperso +
    walksperso + soserrors + sbsobp, data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max 
-1875.33  -436.32     5.95   317.20  2995.83 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   15.40179  137.77644   0.112 0.911061    
runs          16.67719    6.05523   2.754 0.006218 ** 
hits          -9.36268    3.19657  -2.929 0.003643 ** 
rbis          29.78056    5.21096   5.715 2.50e-08 ***
sos           -9.39420    2.74037  -3.428 0.000687 ***
sbs          -55.40790   32.66691  -1.696 0.090824 .  
freeagent   1296.99110   95.21092  13.622  < 2e-16 ***
arbitration  862.99943  112.54532   7.668 2.09e-13 ***
runsperso   -245.94935  147.79397  -1.664 0.097058 .  
hitsperso    334.52758  127.03055   2.633 0.008860 ** 
hrsperso     885.20889  489.74386   1.807 0.071618 .  
rbisperso   -646.35105  281.02310  -2.300 0.022088 *  
walksperso  -227.46388  145.33684  -1.565 0.118546    
soserrors     -0.12112    0.08719  -1.389 0.165762    
sbsobp       191.63493   92.97141   2.061 0.040084 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.1 on 322 degrees of freedom
Multiple R-squared:  0.7057,	Adjusted R-squared:  0.6929 
F-statistic: 55.16 on 14 and 322 DF,  p-value: < 2.2e-16

> extractAIC(regfit.bwd)
[1]   15.000 4417.591
```

Both stepwise selection



```
> summary(regfit.both)

Call:
lm(formula = salary ~ runs + hits + rbis + sos + sbs + freeagent +
    arbitration + runsperso + hitsperso + hrsperso + rbisperso +
    walksperso + soserrors + sbsobp, data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-1875.33  -436.32     5.95   317.20  2995.83

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.40179  137.77644   0.112 0.911061
runs          16.67719    6.05523   2.754 0.006218 **
hits          -9.36268    3.19657  -2.929 0.003643 **
rbis          29.78056    5.21096   5.715 2.50e-08 ***
sos           -9.39420    2.74037  -3.428 0.000687 ***
sbs          -55.40790   32.66691  -1.696 0.090824 .
freeagent   1296.99110   95.21092  13.622  < 2e-16 ***
arbitration  862.99943  112.54532   7.668 2.09e-13 ***
runsperso   -245.94935  147.79397  -1.664 0.097058 .
hitsperso    334.52758  127.03055   2.633 0.008860 **
hrsperso     885.20889  489.74386   1.807 0.071618 .
rbisperso   -646.35105  281.02310  -2.300 0.022088 *
walksperso  -227.46388  145.33684  -1.565 0.118546
soserrors     -0.12112    0.08719  -1.389 0.165762
sbsobp       191.63493   92.97141   2.061 0.040084 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 687.1 on 322 degrees of freedom
Multiple R-squared:  0.7057,	Adjusted R-squared:  0.6929
F-statistic: 55.16 on 14 and 322 DF,  p-value: < 2.2e-16

> extractAIC(regfit.both)
[1]   15.000 4417.591
```

Code appendix

```
# 3

library(leaps)

baseball <- read.table("/Users/ssugi/Downloads/baseball.dat.txt", header = TRUE)

head(baseball)

dim(baseball)

summary(baseball)


## all possible subsets

regfit.full=regsubsets(salary~.,data=baseball,nvmax=27)

reg.summary=summary(regfit.full)

reg.summary


which.min(reg.summary$cp) #AIC

coef(regfit.full,8)


mylm <- lm(salary ~ homeruns + rbis + walks + sos + freeagent + arbitration + walksperso + sbsobp, baseball)

summary(mylm)

extractAIC(mylm)


## Forward, Backward and Both Stepwise Selection


intercept_only <- lm(salary ~ 1, data=baseball)

all <- lm(salary ~ ., data=baseball)

regfit.fwd <- step(intercept_only, direction='forward', scope=as.formula(all), trace=FALSE)

summary(regfit.fwd)

extractAIC(regfit.fwd)


regfit.bwd=step(lm(salary~.,data=baseball),direction="backward")

summary(regfit.bwd)

extractAIC(regfit.bwd)


regfit.both = step(lm(salary~.,data=baseball),direction="both")

summary(regfit.both)

extractAIC(regfit.both)
```