hw8

Xinrui Hu

2023-04-11

```
library(MASS)
pros = read.table("http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data")
```

1.

```
df <- pros
df$lbph <- NULL
```

2.

```
head(df)
```

```
##       lcavol  lweight age svi       lcp gleason pgg45       lpsa train
## 1 -0.5798185 2.769459  50   0 -1.386294       6     0 -0.4307829  TRUE
## 2 -0.9942523 3.319626  58   0 -1.386294       6     0 -0.1625189  TRUE
## 3 -0.5108256 2.691243  74   0 -1.386294       7    20 -0.1625189  TRUE
## 4 -1.2039728 3.282789  58   0 -1.386294       6     0 -0.1625189  TRUE
## 5  0.7514161 3.432373  62   0 -1.386294       6     0  0.3715636  TRUE
## 6 -1.0498221 3.228826  50   0 -1.386294       6     0  0.7654678  TRUE
```

3.

```
summary(df)
```

```
##      lcavol           lweight           age             svi        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :0.0000  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:0.0000  
##  Median : 1.4469   Median :3.623   Median :65.00   Median :0.0000  
##  Mean   : 1.3500   Mean   :3.629   Mean   :63.87   Mean   :0.2165  
##  3rd Qu.: 2.1270   3rd Qu.:3.876   3rd Qu.:68.00   3rd Qu.:0.0000  
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   :1.0000  
##       lcp             gleason          pgg45            lpsa        
##  Min.   :-1.3863   Min.   :6.000   Min.   :  0.00   Min.   :-0.4308  
##  1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00   1st Qu.: 1.7317  
##  Median :-0.7985   Median :7.000   Median : 15.00   Median : 2.5915  
##  Mean   :-0.1794   Mean   :6.753   Mean   : 24.38   Mean   : 2.4784  
##  3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00   3rd Qu.: 3.0564  
##  Max.   : 2.9042   Max.   :9.000   Max.   :100.00   Max.   : 5.5829  
##    train        
##  Mode :logical  
##  FALSE:30       
##  TRUE :67       
##                 
##                 
##
```

```
# No missing value
```

4.

```
nrow(df)
```

```
## [1] 97
```

5.

```
lmod <- lm(lpsa ~ ., df)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66408 -0.40279  0.01298  0.39979  1.54622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.575987   1.271683  -0.453 0.651712
## lcavol       0.551524   0.088913   6.203 1.76e-08 ***
## lweight      0.754154   0.187669   4.019 0.000123 ***
## age         -0.016791   0.011161  -1.504 0.136062
## svi          0.702389   0.242432   2.897 0.004750 **
## lcp         -0.102934   0.091538  -1.124 0.263863
## gleason      0.054982   0.160495   0.343 0.732734
## pgg45        0.004759   0.004525   1.052 0.295789
## trainTRUE   -0.018012   0.163825  -0.110 0.912703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 88 degrees of freedom
## Multiple R-squared:  0.6528, Adjusted R-squared:  0.6212
## F-statistic: 20.68 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
lmod <- update(lmod, . ~ . - train)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + svi + lcp + gleason +
##     pgg45, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67014 -0.40402  0.01034  0.39694  1.55244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.598941   1.247446  -0.480 0.632310
## lcavol       0.551786   0.088386   6.243 1.43e-08 ***
## lweight      0.755695   0.186103   4.061 0.000105 ***
## age         -0.017023   0.010899  -1.562 0.121881
```

```
## svi          0.701146   0.240821   2.911 0.004545 **
## lcp         -0.102127   0.090736  -1.126 0.263386
## gleason      0.058274   0.156799   0.372 0.711035
## pgg45        0.004659   0.004407   1.057 0.293324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7065 on 89 degrees of freedom
## Multiple R-squared:  0.6527, Adjusted R-squared:  0.6254
## F-statistic:  23.9 on 7 and 89 DF,  p-value: < 2.2e-16
```

```
lmod <- update(lmod, . ~ . - gleason)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + svi + lcp + pgg45,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66643 -0.43289 -0.00535  0.37290  1.58094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.234573   0.767619  -0.306 0.760626
## lcavol       0.557856   0.086447   6.453 5.39e-09 ***
## lweight      0.747736   0.183980   4.064 0.000103 ***
## age         -0.016593   0.010786  -1.538 0.127459
## svi          0.689677   0.237689   2.902 0.004667 **
## lcp         -0.100802   0.090230  -1.117 0.266898
## pgg45        0.005688   0.003412   1.667 0.099012 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7031 on 90 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.629
## F-statistic: 28.13 on 6 and 90 DF,  p-value: < 2.2e-16
```

```
lmod <- update(lmod, . ~ . - lcp)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + svi + pgg45, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71259 -0.42880 -0.00943  0.42875  1.49167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.194926   0.767843  -0.254 0.800175
## lcavol       0.513814   0.077039   6.670 1.95e-09 ***
## lweight      0.746069   0.184224   4.050 0.000108 ***
```

```
## age         -0.014866   0.010689  -1.391 0.167679
## svi          0.578828   0.216282   2.676 0.008830 **
## pgg45        0.003952   0.003042   1.299 0.197207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7041 on 91 degrees of freedom
## Multiple R-squared:  0.6474, Adjusted R-squared:  0.628
## F-statistic: 33.41 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
lmod <- update(lmod, . ~ . - pgg45)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + svi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77040 -0.44899 -0.01719  0.43825  1.59996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.27369    0.76830  -0.356 0.722482
## lcavol       0.53631    0.07535   7.118 2.35e-10 ***
## lweight      0.72442    0.18415   3.934 0.000162 ***
## age         -0.01166    0.01044  -1.117 0.266939
## svi          0.66422    0.20682   3.212 0.001819 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7067 on 92 degrees of freedom
## Multiple R-squared:  0.6408, Adjusted R-squared:  0.6252
## F-statistic: 41.03 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
lmod <- update(lmod, . ~ . - age)
summary(lmod)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77745 -0.45004 -0.00254  0.44305  1.57574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.77716    0.62300  -1.247 0.215367
## lcavol       0.52585    0.07486   7.024 3.49e-10 ***
## lweight      0.66177    0.17564   3.768 0.000289 ***
## svi          0.66567    0.20709   3.214 0.001798 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7076 on 93 degrees of freedom
## Multiple R-squared:  0.6359, Adjusted R-squared:  0.6242
## F-statistic: 54.15 on 3 and 93 DF,  p-value: < 2.2e-16
```

6.

```
require(leaps)
```

```
## Loading required package: leaps
```

```
model <- regsubsets(lpsa ~ ., df)
rs <- summary(model)
rs$which
```

```
##   (Intercept) lcavol lweight   age   svi   lcp gleason pgg45 trainTRUE
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE   FALSE FALSE     FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE   FALSE FALSE     FALSE
## 3        TRUE   TRUE    TRUE FALSE  TRUE FALSE   FALSE FALSE     FALSE
## 4        TRUE   TRUE    TRUE  TRUE  TRUE FALSE   FALSE FALSE     FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE FALSE   FALSE  TRUE     FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE   FALSE  TRUE     FALSE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE    TRUE  TRUE     FALSE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE    TRUE  TRUE      TRUE
```

```
rs$rss
```

```
## [1] 58.91478 51.74218 46.56844 45.94543 45.10891 44.49192 44.42298 44.41688
```

```
n <- nrow(df)
p <- 2:10
AIC <- n*log(rs$rss / n) + 2 * p
```

```
## Warning in n * log(rs$rss/n) + 2 * p: longer object length is not a multiple of
## shorter object length
```

```
AIC
```

```
## [1] -44.36603 -54.95846 -63.17744 -62.48389 -62.26623 -61.60212 -59.75254
## [8] -57.76587 -28.36603
```

```
model1 <- lm(lpsa ~ lcavol+lweight+svi, df)
summary(model1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77745 -0.45004 -0.00254  0.44305  1.57574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.77716    0.62300  -1.247 0.215367
## lcavol       0.52585    0.07486   7.024 3.49e-10 ***
## lweight      0.66177    0.17564   3.768 0.000289 ***
## svi          0.66567    0.20709   3.214 0.001798 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7076 on 93 degrees of freedom
## Multiple R-squared:  0.6359, Adjusted R-squared:  0.6242
## F-statistic: 54.15 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
# The third model has the lowest value
```

7.

```
BIC <- n*log(rs$rss/n) + p*log(n)
```

```
## Warning in n * log(rs$rss/n) + p * log(n): longer object length is not a
## multiple of shorter object length
```

```
BIC
```

```
## [1] -39.216613 -47.234329 -52.878592 -49.610340 -46.817967 -43.579142 -39.154855
## [8] -34.593467  -2.618925
```

```
# The third model has the lowest value
```

8.

```
rs$adjr2
```

```
## [1] 0.5345839 0.5868977 0.6242063 0.6252038 0.6279840 0.6289953 0.6254081
## [8] 0.6212034
```

```
model2 <- lm(lpsa ~ .-train-gleason, df)
summary(model2)
```

```
##
## Call:
## lm(formula = lpsa ~ . - train - gleason, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66643 -0.43289 -0.00535  0.37290  1.58094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.234573   0.767619  -0.306 0.760626
## lcavol       0.557856   0.086447   6.453 5.39e-09 ***
## lweight      0.747736   0.183980   4.064 0.000103 ***
## age         -0.016593   0.010786  -1.538 0.127459
## svi          0.689677   0.237689   2.902 0.004667 **
## lcp         -0.100802   0.090230  -1.117 0.266898
## pgg45        0.005688   0.003412   1.667 0.099012 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7031 on 90 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.629
## F-statistic: 28.13 on 6 and 90 DF,  p-value: < 2.2e-16
```

```
# The sixth model has the highest value
```