Variable-Selection.R

patriciamaya

2020-12-05

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.2
```

```
data("Carseats")
attach(Carseats)

set.seed(256)
indx <- sample(2,nrow(Carseats), replace=T, prob = c(0.8, 0.2))
train <- Carseats[indx ==1, ]
test <- Carseats[indx ==2, ]

#****VARIABLE SELECTION****
#In general, we would like to reduce num of variables in our regression model.
#Eg: forward, backward, and stepwise selection.

#Extreme cases
full <- lm(Sales ~ . , data = train)
null <- lm(Sales ~ 1 , data = train) #only considers intercept, no variables as inputs

#**FORWARD SELECTION**
#Considers one variable at a time:
  #if it improves the model, we include the variable
  #otherwise, we don't include variable.
#We can look at r-squared, f-test, or AIC.

step(null, scope = list(lower=null, upper =full), direction = "forward")
```

```
## Start:  AIC=704.82
## Sales ~ 1
##
##               Df Sum of Sq    RSS    AIC
## + ShelveLoc    2    933.40 1779.0 566.68
## + Price        1    512.52 2199.9 636.24
## + Advertising  1    224.93 2487.5 677.65
## + Age          1    151.69 2560.7 687.43
## + US           1    109.35 2603.1 692.95
## + Income       1     37.03 2675.4 702.19
## + CompPrice    1     20.26 2692.2 704.29
## <none>                     2712.4 704.82
## + Education    1      6.42 2706.0 706.02
## + Population   1      1.98 2710.4 706.57
## + Urban        1      0.73 2711.7 706.73
```

```
##
## Step:  AIC=566.68
## Sales ~ ShelveLoc
##
##               Df Sum of Sq    RSS    AIC
## + Price        1    556.78 1222.3 442.18
## + Age          1    169.00 1610.0 535.04
## + Advertising  1    168.88 1610.2 535.07
## + US           1     70.67 1708.4 555.02
## + Income       1     57.61 1721.4 557.59
## <none>                     1779.0 566.68
## + CompPrice    1      9.98 1769.0 566.78
## + Education    1      7.20 1771.8 567.32
## + Population   1      4.65 1774.4 567.80
## + Urban        1      1.55 1777.5 568.39
##
## Step:  AIC=442.18
## Sales ~ ShelveLoc + Price
##
##               Df Sum of Sq     RSS    AIC
## + CompPrice    1    400.36  821.89 310.44
## + Age          1    229.61  992.64 374.06
## + Advertising  1    187.36 1034.89 388.10
## + US           1     80.39 1141.87 421.25
## + Income       1     59.31 1162.94 427.42
## <none>                     1222.26 442.18
## + Urban        1      6.51 1215.75 442.38
## + Population   1      4.03 1218.23 443.07
## + Education    1      1.73 1220.53 443.70
##
## Step:  AIC=310.44
## Sales ~ ShelveLoc + Price + CompPrice
##
##               Df Sum of Sq    RSS    AIC
## + Advertising  1   219.698 602.20 207.63
## + Age          1   201.540 620.35 217.64
## + Income       1    82.816 739.08 276.65
## + US           1    82.354 739.54 276.86
## + Population   1    29.563 792.33 300.10
## <none>                     821.89 310.44
## + Education    1     1.541 820.35 311.81
## + Urban        1     1.197 820.70 311.95
##
## Step:  AIC=207.63
## Sales ~ ShelveLoc + Price + CompPrice + Advertising
##
##              Df Sum of Sq    RSS    AIC
## + Age         1   189.653 412.54  82.16
## + Income      1    64.017 538.18 171.75
## <none>                    602.20 207.63
## + Population  1     1.801 600.39 208.62
## + US          1     1.284 600.91 208.91
## + Education   1     0.918 601.28 209.11
## + Urban       1     0.082 602.11 209.58
```

```
##
## Step:  AIC=82.16
## Sales ~ ShelveLoc + Price + CompPrice + Advertising + Age
##
##              Df Sum of Sq    RSS    AIC
## + Income      1    60.846 351.70 30.385
## <none>                    412.54 82.160
## + Urban       1     1.387 411.16 83.026
## + Education   1     0.954 411.59 83.380
## + US          1     0.513 412.03 83.741
## + Population  1     0.277 412.27 83.934
##
## Step:  AIC=30.39
## Sales ~ ShelveLoc + Price + CompPrice + Advertising + Age + Income
##
##              Df Sum of Sq    RSS    AIC
## <none>                    351.70 30.385
## + US          1   1.71959 349.98 30.733
## + Urban       1   0.57903 351.12 31.830
## + Population  1   0.55202 351.14 31.856
## + Education   1   0.24810 351.45 32.147
```

```
##
## Call:
## lm(formula = Sales ~ ShelveLoc + Price + CompPrice + Advertising +
##     Age + Income, data = train)
##
## Coefficients:
##     (Intercept)    ShelveLocGood  ShelveLocMedium            Price
##         5.69723          4.84599          1.94907         -0.09483
##       CompPrice      Advertising              Age           Income
##         0.09065          0.11581         -0.04630          0.01537
```

```
#check all possibilities from null case to full case

#Start: AIC=704.82
  #Sales ~ 1
  #if I add X variable, see how AIC decreases
  #we want the LOWEST AIC
    #we add ShelveLoc
#Step:  AIC=566.68
  #Sales ~ ShelveLoc
    #now we add Price as it reduces AIC
#... repeat process until AIC is *NOT* reduced.
#first row is <none>

#final model display at the end

#**BACKWARD ELIMINATION**
step(full, scope = list(lower=null, upper =full), direction = "backward")
```

```
## Start:  AIC=35.51
## Sales ~ CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
##
```

```
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.21  348.92  33.71
## - Education    1      0.32  349.03  33.82
## - Urban        1      0.71  349.42  34.20
## - US           1      1.62  350.33  35.07
## <none>                      348.71  35.51
## - Income       1     60.63  409.34  87.53
## - Advertising  1    106.78  455.49 123.53
## - Age          1    184.68  533.39 176.74
## - CompPrice    1    411.16  759.87 296.00
## - ShelveLoc    2    914.55 1263.26 465.30
## - Price        1   1042.01 1390.72 499.69
##
## Step:  AIC=33.71
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Education    1      0.39  349.31  32.09
## - Urban        1      0.68  349.60  32.37
## - US           1      1.95  350.86  33.58
## <none>                      348.92  33.71
## - Income       1     60.52  409.44  85.61
## - Advertising  1    124.41  473.32 134.48
## - Age          1    185.84  534.76 175.60
## - CompPrice    1    418.37  767.28 297.27
## - ShelveLoc    2    914.78 1263.70 463.42
## - Price        1   1046.52 1395.44 498.84
##
## Step:  AIC=32.09
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Urban        1      0.67  349.98  30.73
## - US           1      1.81  351.12  31.83
## <none>                      349.31  32.09
## - Income       1     61.25  410.56  84.54
## - Advertising  1    124.02  473.33 132.48
## - Age          1    185.83  535.14 173.85
## - CompPrice    1    418.52  767.83 295.52
## - ShelveLoc    2    914.55 1263.86 461.46
## - Price        1   1051.41 1400.72 498.11
##
## Step:  AIC=30.73
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + US
##
##               Df Sum of Sq     RSS    AIC
## - US           1      1.72  351.70  30.39
## <none>                      349.98  30.73
## - Income       1     62.05  412.03  83.74
## - Advertising  1    124.24  474.21 131.11
## - Age          1    185.17  535.15 171.85
```

```
## - CompPrice    1    423.97  773.95 296.19
## - ShelveLoc    2    916.60 1266.58 460.19
## - Price        1   1051.26 1401.24 496.23
##
## Step:  AIC=30.39
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                      351.70  30.39
## - Income       1     60.85  412.54  82.16
## - Age          1    186.48  538.18 171.75
## - Advertising  1    190.10  541.80 174.01
## - CompPrice    1    422.34  774.04 294.23
## - ShelveLoc    2    915.03 1266.73 458.23
## - Price        1   1049.93 1401.63 494.33
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = train)
##
## Coefficients:
##     (Intercept)        CompPrice           Income      Advertising
##         5.69723          0.09065          0.01537          0.11581
##           Price    ShelveLocGood  ShelveLocMedium              Age
##        -0.09483          4.84599          1.94907         -0.04630
```

```
#check all possibilities from full case to null case


#**STEPWISE ELIMINATION**
step(full, scope = list(lower=null, upper =full), direction = "both")
```

```
## Start:  AIC=35.51
## Sales ~ CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.21  348.92  33.71
## - Education    1      0.32  349.03  33.82
## - Urban        1      0.71  349.42  34.20
## - US           1      1.62  350.33  35.07
## <none>                      348.71  35.51
## - Income       1     60.63  409.34  87.53
## - Advertising  1    106.78  455.49 123.53
## - Age          1    184.68  533.39 176.74
## - CompPrice    1    411.16  759.87 296.00
## - ShelveLoc    2    914.55 1263.26 465.30
## - Price        1   1042.01 1390.72 499.69
##
## Step:  AIC=33.71
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + Urban + US
##
```

```
##               Df Sum of Sq     RSS    AIC
## - Education    1      0.39  349.31  32.09
## - Urban        1      0.68  349.60  32.37
## - US           1      1.95  350.86  33.58
## <none>                      348.92  33.71
## + Population   1      0.21  348.71  35.51
## - Income       1     60.52  409.44  85.61
## - Advertising  1    124.41  473.32 134.48
## - Age          1    185.84  534.76 175.60
## - CompPrice    1    418.37  767.28 297.27
## - ShelveLoc    2    914.78 1263.70 463.42
## - Price        1   1046.52 1395.44 498.84
##
## Step:  AIC=32.09
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Urban        1      0.67  349.98  30.73
## - US           1      1.81  351.12  31.83
## <none>                      349.31  32.09
## + Education    1      0.39  348.92  33.71
## + Population   1      0.28  349.03  33.82
## - Income       1     61.25  410.56  84.54
## - Advertising  1    124.02  473.33 132.48
## - Age          1    185.83  535.14 173.85
## - CompPrice    1    418.52  767.83 295.52
## - ShelveLoc    2    914.55 1263.86 461.46
## - Price        1   1051.41 1400.72 498.11
##
## Step:  AIC=30.73
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + US
##
##               Df Sum of Sq     RSS    AIC
## - US           1      1.72  351.70  30.39
## <none>                      349.98  30.73
## + Urban        1      0.67  349.31  32.09
## + Education    1      0.38  349.60  32.37
## + Population   1      0.25  349.73  32.49
## - Income       1     62.05  412.03  83.74
## - Advertising  1    124.24  474.21 131.11
## - Age          1    185.17  535.15 171.85
## - CompPrice    1    423.97  773.95 296.19
## - ShelveLoc    2    916.60 1266.58 460.19
## - Price        1   1051.26 1401.24 496.23
##
## Step:  AIC=30.39
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                      351.70  30.39
## + US           1      1.72  349.98  30.73
```

```
## + Urban        1      0.58  351.12  31.83
## + Population   1      0.55  351.14  31.86
## + Education    1      0.25  351.45  32.15
## - Income       1     60.85  412.54  82.16
## - Age          1    186.48  538.18 171.75
## - Advertising  1    190.10  541.80 174.01
## - CompPrice    1    422.34  774.04 294.23
## - ShelveLoc    2    915.03 1266.73 458.23
## - Price        1   1049.93 1401.63 494.33
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = train)
##
## Coefficients:
##     (Intercept)        CompPrice           Income      Advertising
##         5.69723          0.09065          0.01537          0.11581
##           Price    ShelveLocGood  ShelveLocMedium              Age
##        -0.09483          4.84599          1.94907         -0.04630
```