Prediction Project

Angelicaqj

5/10/2021

Predicting crime rates in Boston data. The Boston data set is in the MASS package, and first you will need to load it.

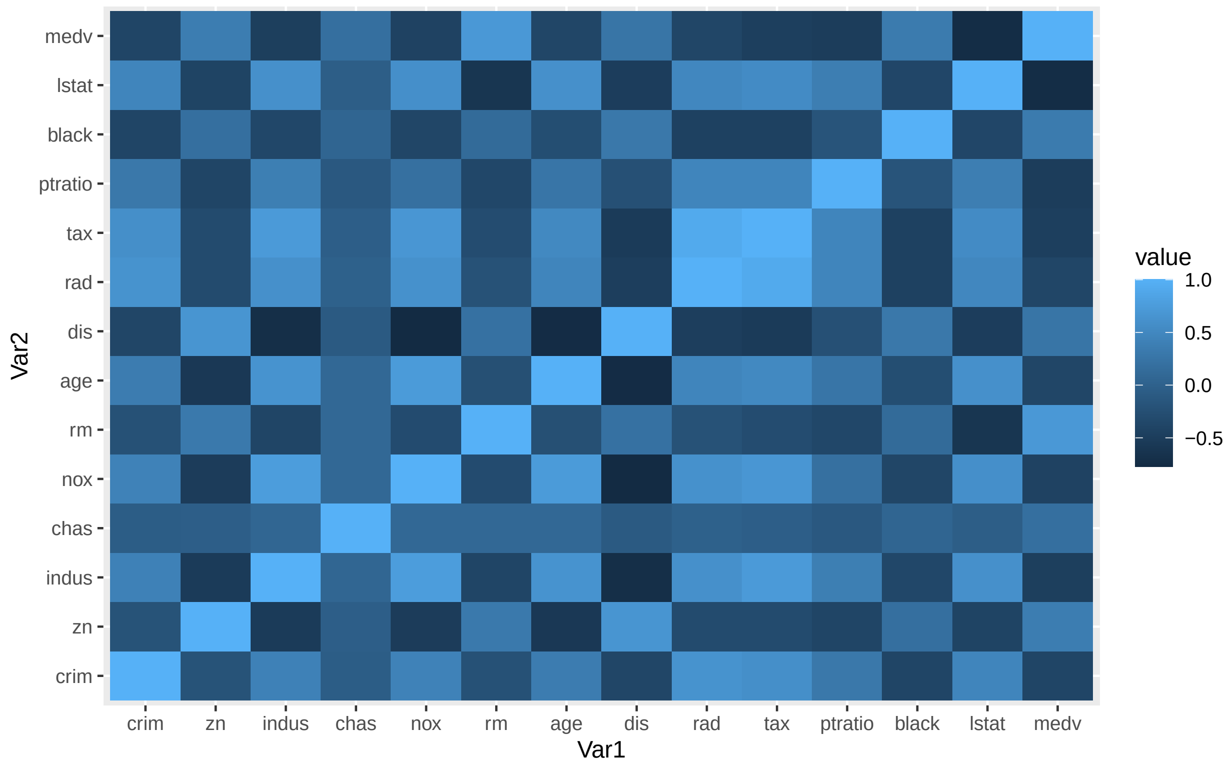
```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

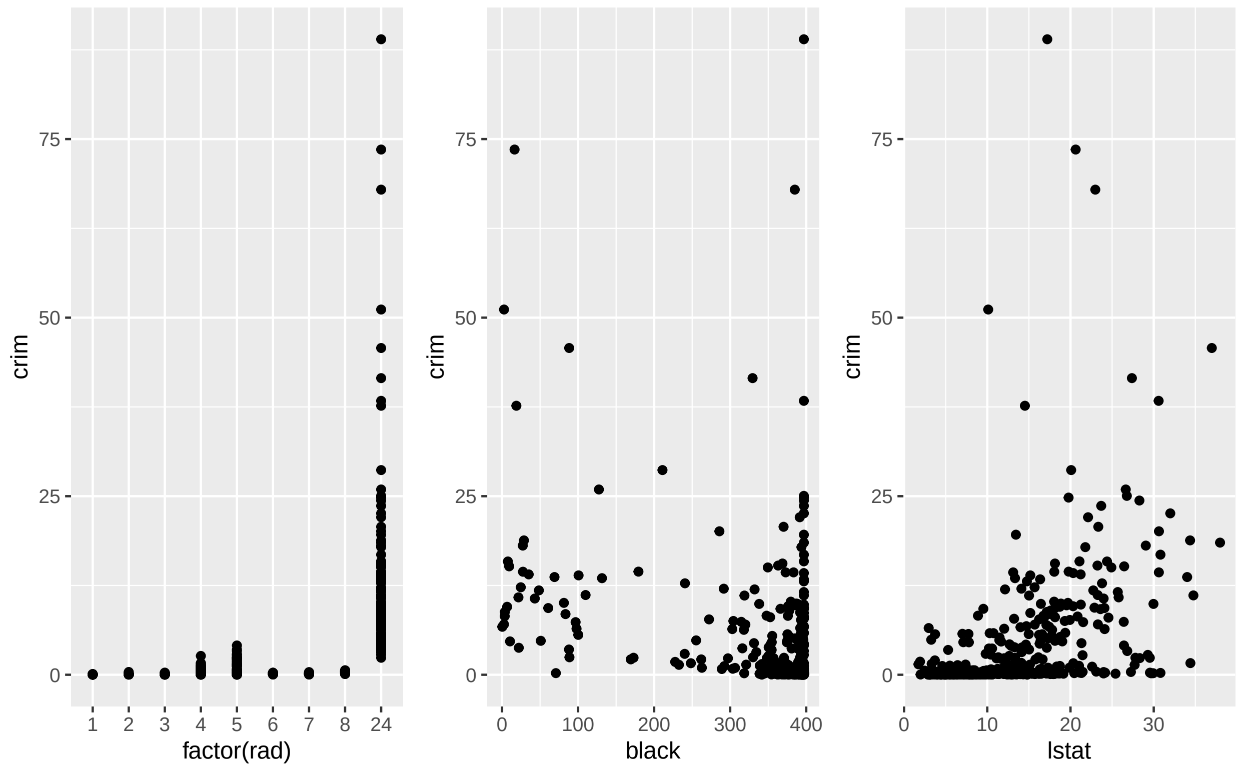
Build a regression model to predict the crime rate (crim) in Boston suburbs based on the other provided variables.

The solution includes:

- A brief exploratory analysis.
- A description of the set of regression models to be considered.
- A description of how the models were evaluated.
- A summary of one model, based on the analysis, is the best among to be considered.

Exploratory Analysis





Generalize a linear model

```
##
## Call:
## glm(formula = crim ~ ., data = Boston)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.924  -2.120  -0.353   1.019  75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 41.46327)
##
##     Null deviance: 37363  on 505  degrees of freedom
## Residual deviance: 20400  on 492  degrees of freedom
## AIC: 3336.5
##
## Number of Fisher Scoring iterations: 2
```

The approach that we have been taking in this dataset is to use regression as a way of summarizing relationships between some of the variables in the dataset.

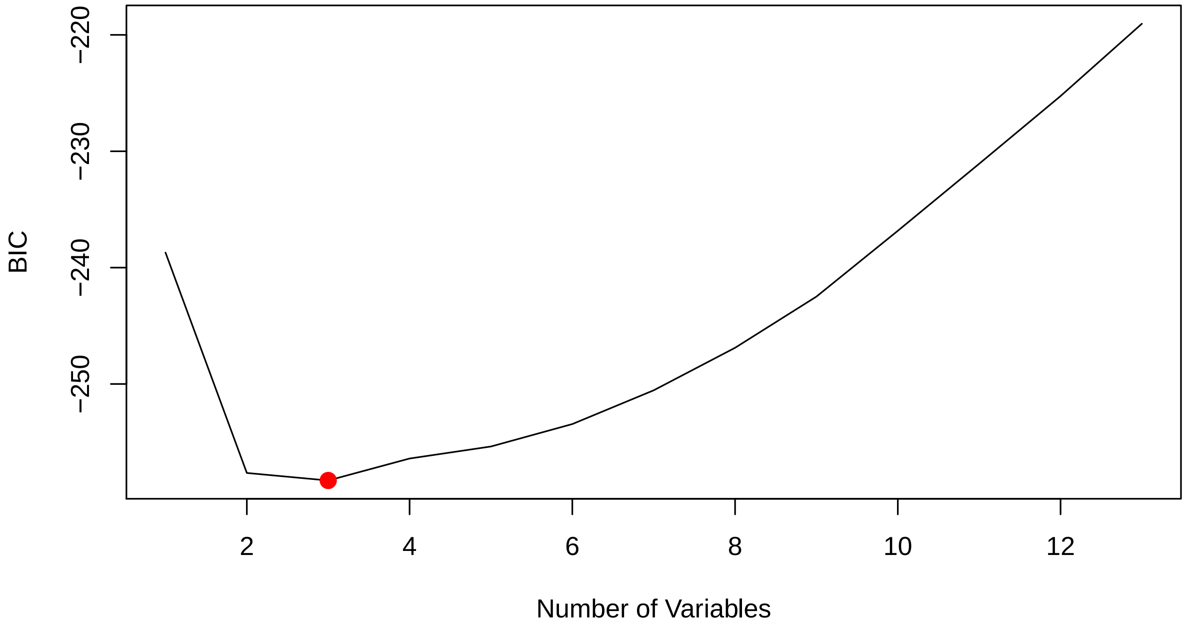
Set Regression Models and evaluation

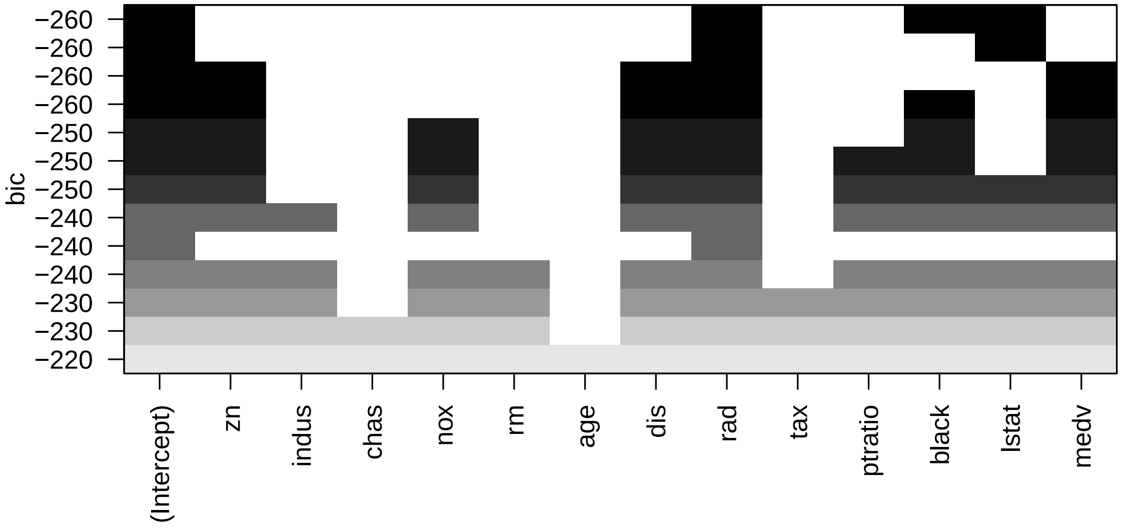
Best subset selection by identifying the best model that contain a given number of predictors.

```
##  [1] 0.3912567 0.4207965 0.4286123 0.4334892 0.4392738 0.4440173 0.4476594
##  [8] 0.4504606 0.4524408 0.4530572 0.4535605 0.4540031 0.4540104
```

It seems that the R^2 statistic increases from 39%, when only one variable is included in the model, to 45%, when all variables are included. As expected, the R^2 statistic increases monotonically as more variables are included.

```
## [1] 3
```





In the first plot, we see that there are three variables that share a BIC close to -260. These three variables are representing in the second plot as `rad`, `black`, and `lstat`, that contain the lowest BIC.

```
## Subset selection object
## Call: regsubsets.formula(crim ~ ., data = Boston, nvmax = 13, method = "forward")
## 13 Variables  (and intercept)
##         Forced in Forced out
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## medv        FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: forward
##           zn  indus chas nox rm  age dis rad tax ptratio black lstat medv
## 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 ) "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
```

```
## Subset selection object
## Call: regsubsets.formula(crim ~ ., data = Boston, nvmax = 13, method = "backward")
## 13 Variables  (and intercept)
##         Forced in Forced out
## zn          FALSE      FALSE
## indus       FALSE      FALSE
## chas        FALSE      FALSE
## nox         FALSE      FALSE
## rm          FALSE      FALSE
## age         FALSE      FALSE
## dis         FALSE      FALSE
## rad         FALSE      FALSE
## tax         FALSE      FALSE
## ptratio     FALSE      FALSE
## black       FALSE      FALSE
## lstat       FALSE      FALSE
## medv        FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: backward
##           zn  indus chas nox rm  age dis rad tax ptratio black lstat medv
## 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 ) "*" "*"   "*"  "*" "*" "*" "*" "*" "*" "*"     "*"   "*"   "*"
```

We can see that using forward and backward stepwise selection, the best one-variable model contains only `rad`.

```
##  (Intercept)          rad        black        lstat 
## -0.372585457  0.488172386 -0.009471639  0.213595700
```

```
##  (Intercept)          rad        black        lstat 
## -0.372585457  0.488172386 -0.009471639  0.213595700
```

```
## (Intercept)         dis         rad        medv 
##   3.4931998  -0.3241247   0.5152751  -0.1584437
```

By looking at the coefficients for these three selections, the best one-variable through three-variable models are each identical for best subset and forward selection.

```
## [1] 43.80667 43.80505
```

Summary

The best model according the Best Subset selection and Forward Stepwise selection is a model with three variables. The BIC plot shows noticeable that `rad`, `black`, and `lstat` have the lowest BIC. Concluding, the regression model to predict the crime rate (crim) in Boston suburbs is the following:

$$crime_i = \beta_0 + \beta_1 rad + \beta_2 black + \beta_3 lstat + \epsilon_i$$

Based on the coefficients the final prediction model is the following:

$$crime_i = -0.37 + 0.49rad - 0.01black + 0.21lstat$$

```
## 
## Call:
## lm(formula = crim ~ rad + black + lstat, data = Boston)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -11.023  -1.713  -0.281   0.873  77.716
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.372585   1.641557  -0.227  0.82054
## rad          0.488172   0.040422  12.077  < 2e-16 ***
## black       -0.009472   0.003615  -2.620  0.00905 **
## lstat        0.213596   0.047447   4.502 8.39e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.521 on 502 degrees of freedom
## Multiple R-squared:  0.4286, Adjusted R-squared:  0.4252
## F-statistic: 125.5 on 3 and 502 DF,  p-value: < 2.2e-16
```