# MRR_TP2

*Zuo You &Hui Jingzhuo*

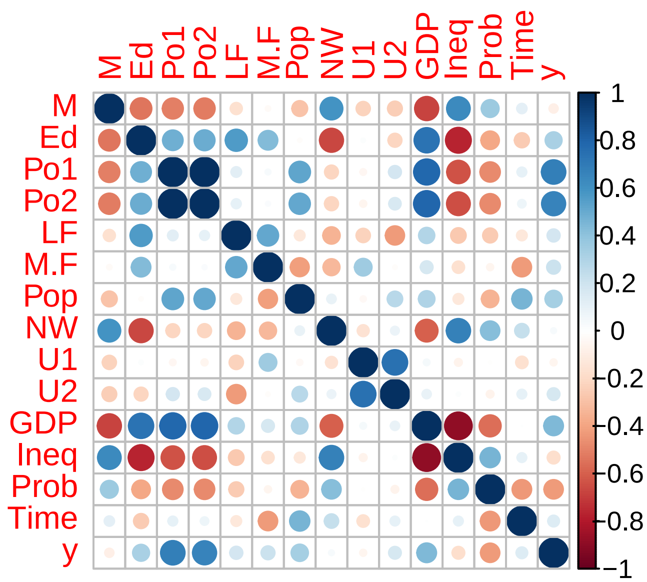
*2019/10/18*

## Application THE Boston housing data set

### (a) onload the data

At the beginning, we took a fisrt sight of our data set:

Here we found one quantitive variable which is So, indicator variable for a Southern state. We can not conclude this kind of variable to our regression model so we just deleted it from the table.

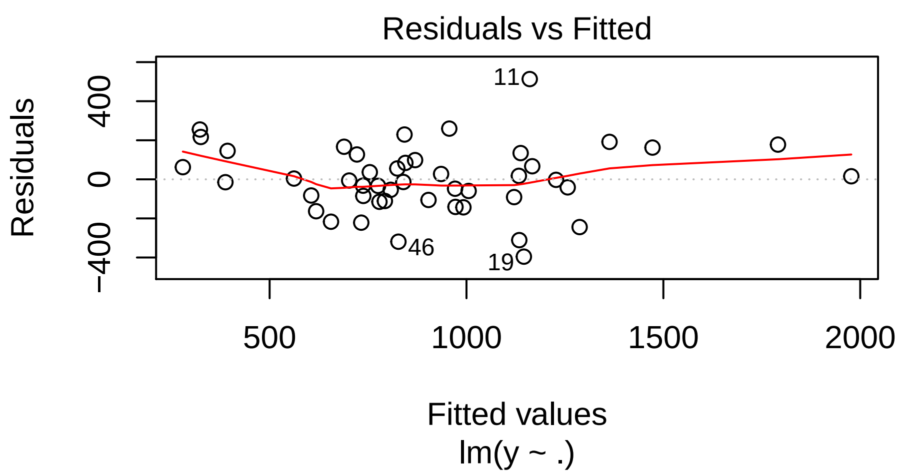


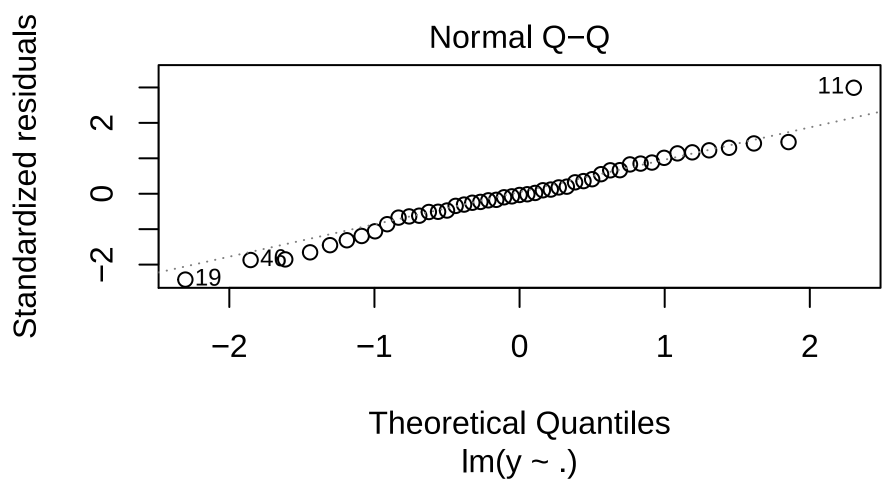

From the corrplot above we can see that some of the variables have a very high level of linear relations, like Po1 and Po2, U1 and U2, GDP and Ineq.

For the first model, we try to build a multiple regression with all 14 the variables to explain the target:

```
##
## Call:
## lm(formula = y ~ ., data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.72  -98.25   -6.12  112.90  513.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5980.6789  1596.6590  -3.746 0.000711 ***
## M               8.7729     4.0872   2.146 0.039520 *
## Ed             18.8230     6.1003   3.086 0.004170 **
## Po1            19.2730    10.4401   1.846 0.074152 .
## Po2           -10.9217    11.5358  -0.947 0.350855
## LF             -0.6461     1.2747  -0.507 0.615736
```

```
## M.F             1.7326     1.9792   0.875 0.387876
## Pop            -0.7331     1.2693  -0.578 0.567573
## NW              0.4135     0.5786   0.715 0.480002
## U1             -5.7863     3.8345  -1.509 0.141106
## U2             16.7333     7.9023   2.118 0.042081 *
## GDP             0.9555     0.9928   0.962 0.343041
## Ineq            7.0455     2.0736   3.398 0.001834 **
## Prob        -4863.6294  2213.3168  -2.197 0.035344 *
## Time           -3.4549     6.9912  -0.494 0.624556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205.8 on 32 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7169
## F-statistic: 9.322 on 14 and 32 DF,  p-value: 1.118e-07
```




From the results above, we can say that our model is generally good, since the residuals perfectly follow the normal distribution. But in other words, there are only about 6 variables which have a certain high level of significativity, and the R-squared values are not that high, so we think there are mores things to exploit for our model.

We implemented some model selection methods:

From the summaries of our step by step methods, we find that backward and stepwise selection have the exactly same results which choose 8 variables and has $AIC = 503.93$, the forward selection choose 6 variables with $AIC = 504.79$. Besides, all of them do not change so much R-squared values. Here we choose the model which use the backward selection.

```
## [1] 639.3151
```

```
## [1] 640.1661
```

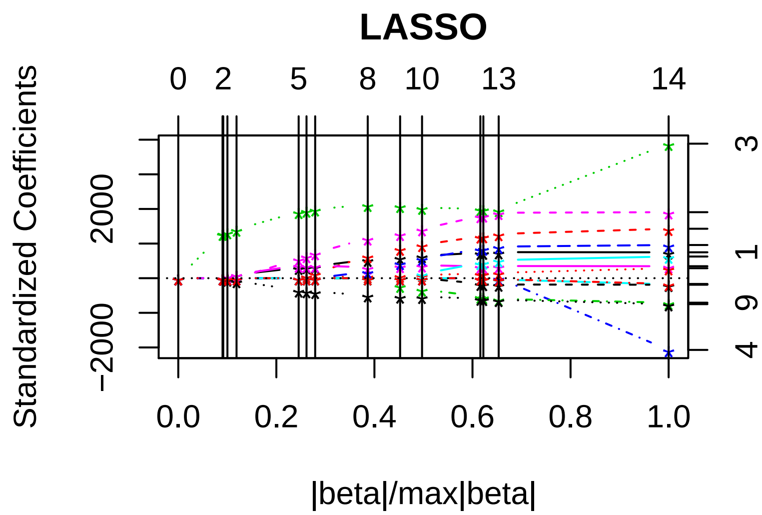
```
## [1] 639.3151
```

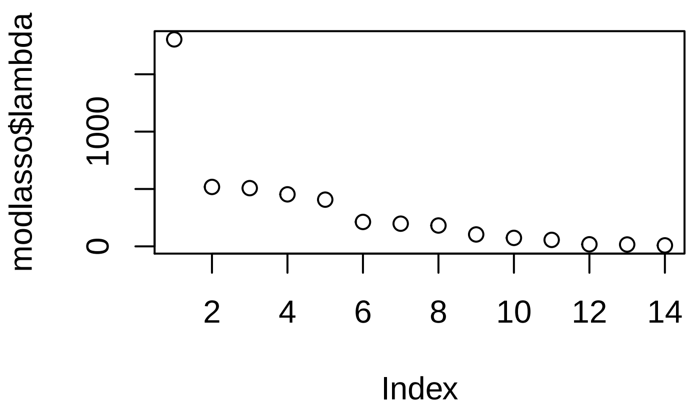
```
##
## Call:
## lm(formula = formula(regbackward), data = UScrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.70 -111.07    3.03  122.15  483.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6426.101   1194.611  -5.379 4.04e-06 ***
## M               9.332      3.350   2.786  0.00828 **
## Ed             18.012      5.275   3.414  0.00153 **
## Po1            10.265      1.552   6.613 8.26e-08 ***
## M.F             2.234      1.360   1.642  0.10874
## U1             -6.087      3.339  -1.823  0.07622 .
## U2             18.735      7.248   2.585  0.01371 *
## Ineq            6.133      1.396   4.394 8.63e-05 ***
## Prob        -3796.032   1490.646  -2.547  0.01505 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195.5 on 38 degrees of freedom
## Multiple R-squared:  0.7888, Adjusted R-squared:  0.7444
## F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

### LASSO

The next step, we try the Lasso regression:

```
## Loaded lars 1.2
```



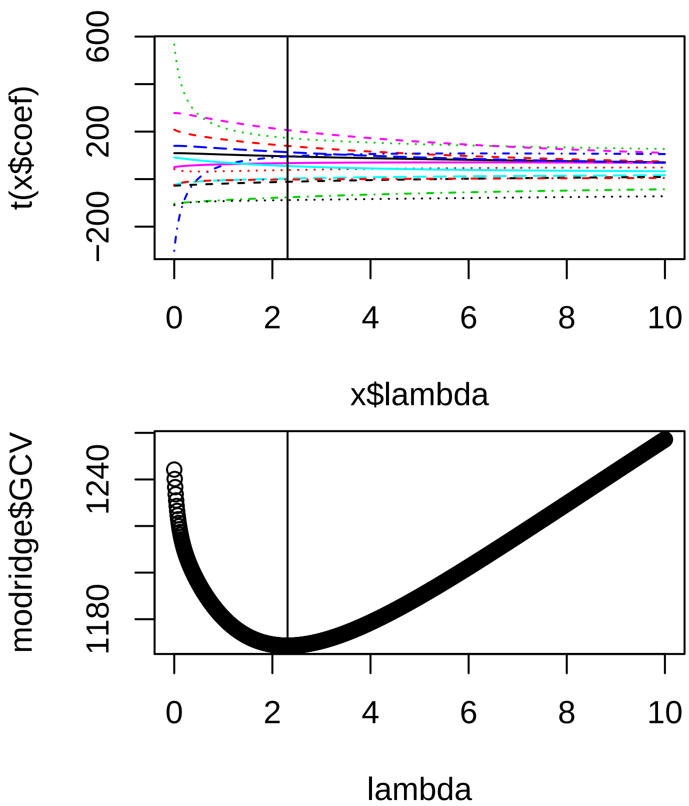

From these two graphs, we can see the evolution of the values of the coefficients for different values of the penalized coefficient. And after the beta bigger than 17, the coefficients become more stable.

```
## [1] 8.573692
```

With the help of criteria RSS, we choose the lambda which is 0.5946502. And we found that the residual standard error is less than the Previous method but the difference is small.

```
##            M            Ed           Po1           Po2            LF
##    8.7764159    18.6853124    18.6151367   -10.1641584    -0.6111188
##          M.F           Pop            NW            U1            U2
##    1.7344270    -0.7311612     0.4002728    -5.7253905    16.6774157
##          GDP          Ineq          Prob          Time
##    0.9446855     7.0411873 -4807.0277108    -3.2574721
```

## RIDGE




For the ridge regression, with the smallest GCV, wo choose the lambda which is 3.23. So we can use the regression model whose lambda equals 3.23.

```
##                           M           Ed          Po1          Po2
## -5.781191e+03  7.708302e+00  1.263330e+01  5.894888e+00  3.438690e+00
##            LF           M.F          Pop           NW           U1
##  5.992781e-02  2.307170e+00 -2.968627e-01  3.758179e-01 -4.289309e+00
##            U2           GDP         Ineq         Prob         Time
##  1.357394e+01  5.771134e-01  5.228858e+00 -3.925451e+03 -2.507089e-01
```

```
## [1] 213.4161
```

So we obtain the result.

What's more, I think about how about it with the new data.

With linear regression

```
## Start:  AIC=387.34
## y ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + GDP +
##     Ineq + Prob + Time
##
##        Df Sum of Sq     RSS    AIC
## - Pop   1         0  950803 385.34
## - Time  1       320  951123 385.35
## - Po2   1     31067  981870 386.47
## - GDP   1     33169  983973 386.54
## <none>               950803 387.34
## - LF    1     70544 1021347 387.84
```

```
## - NW    1     77256 1028059 388.07
## - Po1   1     79853 1030656 388.16
## - M     1     82984 1033787 388.27
## - Prob  1    115270 1066073 389.35
## - M.F   1    116521 1067324 389.39
## - U2    1    118042 1068845 389.44
## - U1    1    118872 1069675 389.46
## - Ed    1    224397 1175200 392.76
## - Ineq  1    248153 1198956 393.46
##
## Step:  AIC=385.34
## y ~ M + Ed + Po1 + Po2 + LF + M.F + NW + U1 + U2 + GDP + Ineq +
##     Prob + Time
##
##        Df Sum of Sq     RSS    AIC
## - Time  1       344  951147 383.35
## - Po2   1     31968  982771 384.50
## - GDP   1     33464  984267 384.55
## <none>               950803 385.34
## - LF    1     72678 1023481 385.92
## - NW    1     77381 1028185 386.08
## - Po1   1     80296 1031099 386.18
## - M     1     83294 1034098 386.28
## - Prob  1    116602 1067406 387.39
## - U2    1    120991 1071794 387.53
## - U1    1    133390 1084193 387.93
## - M.F   1    139985 1090789 388.15
## - Ed    1    229510 1180314 390.91
## - Ineq  1    296742 1247546 392.85
##
## Step:  AIC=383.35
## y ~ M + Ed + Po1 + Po2 + LF + M.F + NW + U1 + U2 + GDP + Ineq +
##     Prob
##
##        Df Sum of Sq     RSS    AIC
## - GDP   1     35834  986982 382.65
## - Po2   1     50122 1001270 383.15
## <none>               951147 383.35
## - LF    1     73934 1025081 383.97
## - NW    1     77039 1028187 384.08
## - M     1     88240 1039388 384.46
## - U2    1    120671 1071819 385.53
## - Po1   1    122749 1073897 385.60
## - U1    1    134034 1085182 385.97
## - M.F   1    166332 1117479 386.99
## - Prob  1    230327 1181474 388.94
## - Ed    1    258631 1209779 389.77
## - Ineq  1    297340 1248488 390.87
##
## Step:  AIC=382.65
## y ~ M + Ed + Po1 + Po2 + LF + M.F + NW + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## - Po2   1     56185 1043166 382.58
```

```
## <none>                986982 382.65
## - LF    1     69342 1056324 383.02
## - M     1     70275 1057257 383.05
## - NW    1     78895 1065876 383.34
## - Po1   1    145332 1132314 385.45
## - U1    1    155783 1142765 385.78
## - U2    1    157157 1144138 385.82
## - M.F   1    171759 1158740 386.26
## - Prob  1    275143 1262124 389.25
## - Ineq  1    284923 1271905 389.52
## - Ed    1    294803 1281784 389.79
##
## Step:  AIC=382.58
## y ~ M + Ed + Po1 + LF + M.F + NW + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## - LF    1     47495 1090661 382.14
## - NW    1     55607 1098773 382.40
## <none>               1043166 382.58
## - M     1     95361 1138527 383.65
## - U1    1    156476 1199642 385.48
## - M.F   1    178890 1222056 386.12
## - U2    1    184512 1227678 386.29
## - Prob  1    248464 1291630 388.06
## - Ed    1    275440 1318606 388.79
## - Ineq  1    328209 1371375 390.16
## - Po1   1    697025 1740191 398.50
##
## Step:  AIC=382.14
## y ~ M + Ed + Po1 + M.F + NW + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## - NW    1     36281 1126942 381.29
## <none>               1090661 382.14
## - U1    1    113372 1204033 383.60
## - M     1    113428 1204089 383.61
## - M.F   1    131979 1222640 384.14
## - U2    1    174414 1265075 385.34
## - Prob  1    208023 1298684 386.25
## - Ed    1    228380 1319041 386.80
## - Ineq  1    318232 1408893 389.10
## - Po1   1    869297 1959957 400.66
##
## Step:  AIC=381.29
## y ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## <none>               1126942 381.29
## - M.F   1    118686 1245628 382.79
## - U1    1    120746 1247688 382.85
## - U2    1    174858 1301800 384.34
## - Prob  1    177705 1304647 384.41
## - M     1    184277 1311219 384.59
## - Ed    1    200828 1327770 385.03
```

```
## - Ineq  1    444635 1571577 390.93
## - Po1   1   1331977 2458919 406.60
```

with the selection of various:

```
##        1
## 386.9595
```

The linear regression backward:

```
##        1
## 303.2173
```

# LASSO

```
## [1] 353.8263
```

# Ridge

For the ridge regression, with the smallest GCV, wo choose the lambda which is 3.23. So we can use the regression model whose lambda equals 3.23.

```
## [1] 355.6539
```

That's all. I find that for these data, the linear regression backward and the lasso regression is better than Ridge regression. And the normal linear regression fit the new data worse.