Homework 8

Anonymous

2024-03-05

Question 11.1

Using the crime data set uscrime.txt from 8.2, 9.1, and 10.1, build a regression model using:

1. Stepwise regression
2. Lasso
3. Elastic net For Parts 2 and 3, remember to scale the data first - otherwise, the regression coefficients will be on different scales and the constraint won't have desired effect.

For part 2 and 3, use the glmnet function in R.

Notes on R:

- For the elastic net model, what we called Lambda in the videos, glmnet calls "alpha"; you can get a range of results by varying alpha from 1 (lasso) to 0 (ridge regression) [and, of course, other values of alpha in between].
- In a function call like glmnet (x, y, family = "gaussian", alpha = 1) the predictors x need to be in R's matrix format, rather than data frame format. You can convert a data frame to a matrix using as.matirix - for example, x <- as.matrix(data[,1:n-1])
- Rather than specifying a value of T, glmnet returns models for a variety of values of T.

  For practice, I decided to run both *forward selection* and *backward elimination* models for the data set and then proceeded with *Stepwise Regression.*

```
crime_data <- read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\uscrime.txt",
                         header = TRUE)

#Performing Backward Elimination
model_back <- lm(Crime~., data = crime_data)
step(model_back,
     direction = "backward")
```

```
## Start:  AIC=514.65
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
```

```
## - So      1        29 1354974 512.65
## - LF      1      8917 1363862 512.96
## - Time    1     10304 1365250 513.00
## - Pop     1     14122 1369068 513.14
## - NW      1     18395 1373341 513.28
## - M.F     1     31967 1386913 513.74
## - Wealth  1     37613 1392558 513.94
## - Po2     1     37919 1392865 513.95
## <none>                1354946 514.65
## - U1      1     83722 1438668 515.47
## - Po1     1    144306 1499252 517.41
## - U2      1    181536 1536482 518.56
## - M       1    193770 1548716 518.93
## - Prob    1    199538 1554484 519.11
## - Ed      1    402117 1757063 524.86
## - Ineq    1    423031 1777977 525.42
##
## Step:  AIC=512.65
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - Time    1     10341 1365315 511.01
## - LF      1     10878 1365852 511.03
## - Pop     1     14127 1369101 511.14
## - NW      1     21626 1376600 511.39
## - M.F     1     32449 1387423 511.76
## - Po2     1     37954 1392929 511.95
## - Wealth  1     39223 1394197 511.99
## <none>                1354974 512.65
## - U1      1     96420 1451395 513.88
## - Po1     1    144302 1499277 515.41
## - U2      1    189859 1544834 516.81
## - M       1    195084 1550059 516.97
## - Prob    1    204463 1559437 517.26
## - Ed      1    403140 1758114 522.89
## - Ineq    1    488834 1843808 525.13
##
## Step:  AIC=511.01
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - LF      1     10533 1375848 509.37
## - NW      1     15482 1380797 509.54
## - Pop     1     21846 1387161 509.75
## - Po2     1     28932 1394247 509.99
## - Wealth  1     36070 1401385 510.23
## - M.F     1     41784 1407099 510.42
## <none>                1365315 511.01
## - U1      1     91420 1456735 512.05
## - Po1     1    134137 1499452 513.41
## - U2      1    184143 1549458 514.95
## - M       1    186110 1551425 515.01
```

```
## - Prob    1    237493 1602808 516.54
## - Ed      1    409448 1774763 521.33
## - Ineq    1    502909 1868224 523.75
##
## Step:  AIC=509.37
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth +
##     Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - NW      1     11675 1387523 507.77
## - Po2     1     21418 1397266 508.09
## - Pop     1     27803 1403651 508.31
## - M.F     1     31252 1407100 508.42
## - Wealth  1     35035 1410883 508.55
## <none>                1375848 509.37
## - U1      1     80954 1456802 510.06
## - Po1     1    123896 1499744 511.42
## - U2      1    190746 1566594 513.47
## - M       1    217716 1593564 514.27
## - Prob    1    226971 1602819 514.54
## - Ed      1    413254 1789103 519.71
## - Ineq    1    500944 1876792 521.96
##
## Step:  AIC=507.77
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Po2     1     16706 1404229 506.33
## - Pop     1     25793 1413315 506.63
## - M.F     1     26785 1414308 506.66
## - Wealth  1     31551 1419073 506.82
## <none>                1387523 507.77
## - U1      1     83881 1471404 508.52
## - Po1     1    118348 1505871 509.61
## - U2      1    201453 1588976 512.14
## - Prob    1    216760 1604282 512.59
## - M       1    309214 1696737 515.22
## - Ed      1    402754 1790276 517.74
## - Ineq    1    589736 1977259 522.41
##
## Step:  AIC=506.33
## Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Pop     1     22345 1426575 505.07
## - Wealth  1     32142 1436371 505.39
## - M.F     1     36808 1441037 505.54
## <none>                1404229 506.33
## - U1      1     86373 1490602 507.13
## - U2      1    205814 1610043 510.76
## - Prob    1    218607 1622836 511.13
## - M       1    307001 1711230 513.62
```

```
## - Ed      1    389502 1793731 515.83
## - Ineq    1    608627 2012856 521.25
## - Po1     1   1050202 2454432 530.57
##
## Step:  AIC=505.07
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - Wealth  1     26493 1453068 503.93
## <none>                1426575 505.07
## - M.F     1     84491 1511065 505.77
## - U1      1     99463 1526037 506.24
## - Prob    1    198571 1625145 509.20
## - U2      1    208880 1635455 509.49
## - M       1    320926 1747501 512.61
## - Ed      1    386773 1813348 514.35
## - Ineq    1    594779 2021354 519.45
## - Po1     1   1127277 2553852 530.44
##
## Step:  AIC=503.93
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##        Df Sum of Sq     RSS    AIC
## <none>              1453068 503.93
## - M.F   1    103159 1556227 505.16
## - U1    1    127044 1580112 505.87
## - Prob  1    247978 1701046 509.34
## - U2    1    255443 1708511 509.55
## - M     1    296790 1749858 510.67
## - Ed    1    445788 1898855 514.51
## - Ineq  1    738244 2191312 521.24
## - Po1   1   1672038 3125105 537.93
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime_data)
##
## Coefficients:
## (Intercept)            M           Ed          Po1          M.F           U1
##    -6426.10        93.32       180.12       102.65        22.34     -6086.63
##          U2         Ineq         Prob
##      187.35        61.33     -3796.03
```

```r
step(model_back,
     direction = "backward", trace = 0)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime_data)
##
## Coefficients:
```

```
## (Intercept)             M            Ed           Po1           M.F            U1
##     -6426.10         93.32        180.12        102.65         22.34      -6086.63
##           U2          Ineq          Prob
##       187.35         61.33      -3796.03
```

```r
#


#Performing Forward Selection
model_forward <- lm(Crime~1, data = crime_data)
step(model_forward,
     scope = formula(lm(Crime~., data = crime_data)),
     direction = "forward")
```

```
## Start:  AIC=561.02
## Crime ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + Po1     1   3253302 3627626 532.94
## + Po2     1   3058626 3822302 535.39
## + Wealth  1   1340152 5540775 552.84
## + Prob    1   1257075 5623853 553.54
## + Pop     1    783660 6097267 557.34
## + Ed      1    717146 6163781 557.85
## + M.F     1    314867 6566061 560.82
## <none>                6880928 561.02
## + LF      1    245446 6635482 561.32
## + Ineq    1    220530 6660397 561.49
## + U2      1    216354 6664573 561.52
## + Time    1    154545 6726383 561.96
## + So      1     56527 6824400 562.64
## + M       1     55084 6825844 562.65
## + U1      1     17533 6863395 562.90
## + NW      1      7312 6873615 562.97
##
## Step:  AIC=532.94
## Crime ~ Po1
##
##          Df Sum of Sq     RSS    AIC
## + Ineq    1    739819 2887807 524.22
## + M       1    616741 3010885 526.18
## + M.F     1    250522 3377104 531.57
## + NW      1    232434 3395192 531.82
## + So      1    219098 3408528 532.01
## + Wealth  1    180872 3446754 532.53
## <none>                3627626 532.94
## + Po2     1    146167 3481459 533.00
## + Prob    1     92278 3535348 533.72
## + LF      1     77479 3550147 533.92
## + Time    1     43185 3584441 534.37
## + U2      1     17848 3609778 534.70
## + Pop     1      5666 3621959 534.86
## + U1      1      2878 3624748 534.90
## + Ed      1       767 3626859 534.93
```

```
##
## Step:  AIC=524.22
## Crime ~ Po1 + Ineq
##
##          Df Sum of Sq     RSS    AIC
## + Ed      1    587050 2300757 515.53
## + M.F     1    454545 2433262 518.17
## + Prob    1    280690 2607117 521.41
## + LF      1    260571 2627236 521.77
## + Wealth  1    213937 2673871 522.60
## + M       1    181236 2706571 523.17
## + Pop     1    130377 2757430 524.04
## <none>                2887807 524.22
## + NW      1     36439 2851369 525.62
## + So      1     33738 2854069 525.66
## + Po2     1     30673 2857134 525.71
## + U1      1      2309 2885498 526.18
## + Time    1       497 2887310 526.21
## + U2      1       253 2887554 526.21
##
## Step:  AIC=515.53
## Crime ~ Po1 + Ineq + Ed
##
##          Df Sum of Sq     RSS    AIC
## + M       1    239405 2061353 512.37
## + Prob    1    234981 2065776 512.47
## + M.F     1    117026 2183731 515.08
## <none>                2300757 515.53
## + Wealth  1     79540 2221218 515.88
## + U2      1     62112 2238646 516.25
## + Time    1     61770 2238987 516.26
## + Po2     1     42584 2258174 516.66
## + Pop     1     39319 2261438 516.72
## + U1      1      7365 2293392 517.38
## + LF      1      7254 2293503 517.39
## + NW      1      4210 2296547 517.45
## + So      1      4135 2296622 517.45
##
## Step:  AIC=512.37
## Crime ~ Po1 + Ineq + Ed + M
##
##          Df Sum of Sq     RSS    AIC
## + Prob    1    258063 1803290 508.08
## + U2      1    200988 1860365 509.55
## + Wealth  1    163378 1897975 510.49
## <none>                2061353 512.37
## + M.F     1     74398 1986955 512.64
## + U1      1     50835 2010518 513.20
## + Po2     1     45392 2015961 513.32
## + Time    1     42746 2018607 513.39
## + NW      1     16488 2044865 513.99
## + Pop     1      8101 2053251 514.19
## + So      1      3189 2058164 514.30
## + LF      1      2988 2058365 514.30
```

```
##
## Step:  AIC=508.08
## Crime ~ Po1 + Ineq + Ed + M + Prob
##
##          Df Sum of Sq     RSS    AIC
## + U2      1    192233 1611057 504.79
## + Wealth  1     86490 1716801 507.77
## + M.F     1     84509 1718781 507.83
## <none>                1803290 508.08
## + U1      1     52313 1750977 508.70
## + Pop     1     47719 1755571 508.82
## + Po2     1     37967 1765323 509.08
## + So      1     21971 1781320 509.51
## + Time    1     10194 1793096 509.82
## + LF      1       990 1802301 510.06
## + NW      1       797 1802493 510.06
##
## Step:  AIC=504.79
## Crime ~ Po1 + Ineq + Ed + M + Prob + U2
##
##          Df Sum of Sq     RSS    AIC
## <none>                1611057 504.79
## + Wealth  1     59910 1551147 505.00
## + U1      1     54830 1556227 505.16
## + Pop     1     51320 1559737 505.26
## + M.F     1     30945 1580112 505.87
## + Po2     1     25017 1586040 506.05
## + So      1     17958 1593098 506.26
## + LF      1     13179 1597878 506.40
## + Time    1      7159 1603898 506.58
## + NW      1       359 1610698 506.78
```

```
##
## Call:
## lm(formula = Crime ~ Po1 + Ineq + Ed + M + Prob + U2, data = crime_data)
##
## Coefficients:
## (Intercept)          Po1         Ineq           Ed            M         Prob
##    -5040.50       115.02        67.65       196.47       105.02     -3801.84
##          U2
##       89.37
```

From Backward Elimination, the final factors are as follows:

1. M
2. Ed
3. Po1
4. M.F
5. U1
6. U2

7. Ineq
8. Prob

And from Forward Selection the following final factors are as follows:

1. M
2. Ed
3. Po1
4. U2
5. Ineq
6. Prob

Part 1 Stepwise Regression Model:

```r
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.3.3
```

```r
#Performing Stepwise Regression
model_stepwise <- lm(Crime~., data = crime_data)
step <- stepAIC(model_stepwise,
     scope = list(lower = formula(lm(Crime~1, data = crime_data)),
                  upper = formula(lm(Crime~., data = crime_data))),
     direction = "both")
```

```
## Start:  AIC=514.65
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - So      1        29 1354974 512.65
## - LF      1      8917 1363862 512.96
## - Time    1     10304 1365250 513.00
## - Pop     1     14122 1369068 513.14
## - NW      1     18395 1373341 513.28
## - M.F     1     31967 1386913 513.74
## - Wealth  1     37613 1392558 513.94
## - Po2     1     37919 1392865 513.95
## <none>                1354946 514.65
## - U1      1     83722 1438668 515.47
## - Po1     1    144306 1499252 517.41
## - U2      1    181536 1536482 518.56
## - M       1    193770 1548716 518.93
## - Prob    1    199538 1554484 519.11
## - Ed      1    402117 1757063 524.86
## - Ineq    1    423031 1777977 525.42
##
```

```
## Step:  AIC=512.65
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - Time    1     10341 1365315 511.01
## - LF      1     10878 1365852 511.03
## - Pop     1     14127 1369101 511.14
## - NW      1     21626 1376600 511.39
## - M.F     1     32449 1387423 511.76
## - Po2     1     37954 1392929 511.95
## - Wealth  1     39223 1394197 511.99
## <none>                1354974 512.65
## - U1      1     96420 1451395 513.88
## + So      1        29 1354946 514.65
## - Po1     1    144302 1499277 515.41
## - U2      1    189859 1544834 516.81
## - M       1    195084 1550059 516.97
## - Prob    1    204463 1559437 517.26
## - Ed      1    403140 1758114 522.89
## - Ineq    1    488834 1843808 525.13
##
## Step:  AIC=511.01
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - LF      1     10533 1375848 509.37
## - NW      1     15482 1380797 509.54
## - Pop     1     21846 1387161 509.75
## - Po2     1     28932 1394247 509.99
## - Wealth  1     36070 1401385 510.23
## - M.F     1     41784 1407099 510.42
## <none>                1365315 511.01
## - U1      1     91420 1456735 512.05
## + Time    1     10341 1354974 512.65
## + So      1        65 1365250 513.00
## - Po1     1    134137 1499452 513.41
## - U2      1    184143 1549458 514.95
## - M       1    186110 1551425 515.01
## - Prob    1    237493 1602808 516.54
## - Ed      1    409448 1774763 521.33
## - Ineq    1    502909 1868224 523.75
##
## Step:  AIC=509.37
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth +
##     Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - NW      1     11675 1387523 507.77
## - Po2     1     21418 1397266 508.09
## - Pop     1     27803 1403651 508.31
## - M.F     1     31252 1407100 508.42
## - Wealth  1     35035 1410883 508.55
```

```
## <none>                 1375848 509.37
## - U1      1     80954 1456802 510.06
## + LF      1     10533 1365315 511.01
## + Time    1      9996 1365852 511.03
## + So      1      3046 1372802 511.26
## - Po1     1    123896 1499744 511.42
## - U2      1    190746 1566594 513.47
## - M       1    217716 1593564 514.27
## - Prob    1    226971 1602819 514.54
## - Ed      1    413254 1789103 519.71
## - Ineq    1    500944 1876792 521.96
##
## Step:  AIC=507.77
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Po2     1     16706 1404229 506.33
## - Pop     1     25793 1413315 506.63
## - M.F     1     26785 1414308 506.66
## - Wealth  1     31551 1419073 506.82
## <none>                 1387523 507.77
## - U1      1     83881 1471404 508.52
## + NW      1     11675 1375848 509.37
## + So      1      7207 1380316 509.52
## + LF      1      6726 1380797 509.54
## + Time    1      4534 1382989 509.61
## - Po1     1    118348 1505871 509.61
## - U2      1    201453 1588976 512.14
## - Prob    1    216760 1604282 512.59
## - M       1    309214 1696737 515.22
## - Ed      1    402754 1790276 517.74
## - Ineq    1    589736 1977259 522.41
##
## Step:  AIC=506.33
## Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Pop     1     22345 1426575 505.07
## - Wealth  1     32142 1436371 505.39
## - M.F     1     36808 1441037 505.54
## <none>                 1404229 506.33
## - U1      1     86373 1490602 507.13
## + Po2     1     16706 1387523 507.77
## + NW      1      6963 1397266 508.09
## + So      1      3807 1400422 508.20
## + LF      1      1986 1402243 508.26
## + Time    1       575 1403654 508.31
## - U2      1    205814 1610043 510.76
## - Prob    1    218607 1622836 511.13
## - M       1    307001 1711230 513.62
## - Ed      1    389502 1793731 515.83
## - Ineq    1    608627 2012856 521.25
```

```
## - Po1     1   1050202 2454432 530.57
##
## Step:  AIC=505.07
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - Wealth  1     26493 1453068 503.93
## <none>                1426575 505.07
## - M.F     1     84491 1511065 505.77
## - U1      1     99463 1526037 506.24
## + Pop     1     22345 1404229 506.33
## + Po2     1     13259 1413315 506.63
## + NW      1      5927 1420648 506.87
## + So      1      5724 1420851 506.88
## + LF      1      5176 1421398 506.90
## + Time    1      3913 1422661 506.94
## - Prob    1    198571 1625145 509.20
## - U2      1    208880 1635455 509.49
## - M       1    320926 1747501 512.61
## - Ed      1    386773 1813348 514.35
## - Ineq    1    594779 2021354 519.45
## - Po1     1   1127277 2553852 530.44
##
## Step:  AIC=503.93
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## <none>                1453068 503.93
## + Wealth  1     26493 1426575 505.07
## - M.F     1    103159 1556227 505.16
## + Pop     1     16697 1436371 505.39
## + Po2     1     14148 1438919 505.47
## + So      1      9329 1443739 505.63
## + LF      1      4374 1448694 505.79
## + NW      1      3799 1449269 505.81
## + Time    1      2293 1450775 505.86
## - U1      1    127044 1580112 505.87
## - Prob    1    247978 1701046 509.34
## - U2      1    255443 1708511 509.55
## - M       1    296790 1749858 510.67
## - Ed      1    445788 1898855 514.51
## - Ineq    1    738244 2191312 521.24
## - Po1     1   1672038 3125105 537.93
```

```
summary(step)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime_data)
##
## 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) -6426.10    1194.61  -5.379 4.04e-06 ***
## M              93.32      33.50   2.786  0.00828 **
## Ed            180.12      52.75   3.414  0.00153 **
## Po1           102.65      15.52   6.613 8.26e-08 ***
## M.F            22.34      13.60   1.642  0.10874
## U1          -6086.63    3339.27  -1.823  0.07622 .
## U2            187.35      72.48   2.585  0.01371 *
## Ineq           61.33      13.96   4.394 8.63e-05 ***
## Prob        -3796.03    1490.65  -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
```

```r
#The summarfy shows M.F and U1 are not significant based on the p-values
#I will remove both and test a new model.

new_model1 <- lm(Crime~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
summary(new_model1)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

Using the stepAIC function in the glmnet library, I built a *Stepwise regression* model. Here the program ran the multiple models to determine the best one based on the stepAIC method.

From running this, these were the factors that were determined to be the best factors by *Stepwise regression*:

1. M
2. Ed
3. Po1
4. M.F
5. U1
6. U2
7. Ineq
8. Prob

Which shows that the Backward Elimination method has similiarity in factors given by the Stepwise Regression method.

When looking at the summary of the model, it showed U1 and M.F were insignificant factors. Therefore I excluded them to test the model further. After I excluded the insignificant factors, which is the purpose of stepwise regression, to make the model simpler, It did not impact the adjusted R^2 's a whole lot. With 8 factors the adjusted R^2 was 73% and the model with 6 factors had an adjusts R^2 of 74%. So, the metrics for both models are close and keeping 8 factors is fine for the model.

Part 2 LASSO

```r
library(glmnet)
```

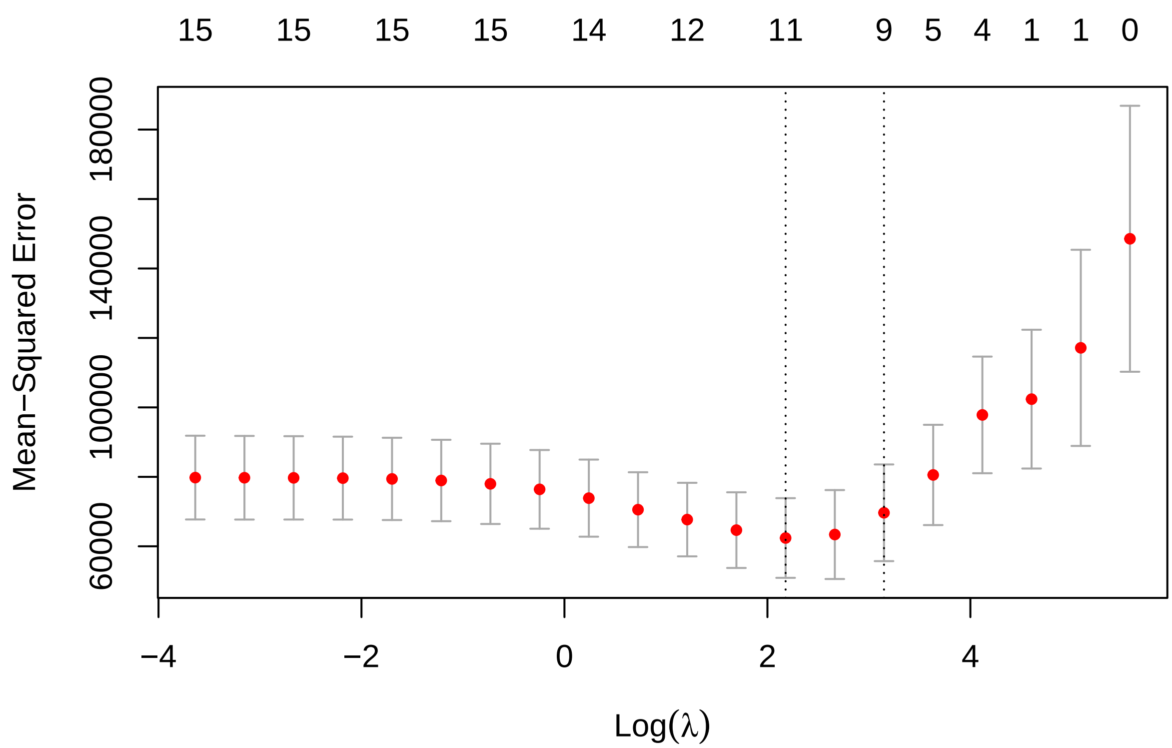
```
## Warning: package 'glmnet' was built under R version 4.3.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.3.3

## Loaded glmnet 4.1-8
```

```r
set.seed(42)
#Performing LASSO
model_lasso <- cv.glmnet(x = as.matrix(crime_data[,-16]),
                         y = as.matrix(crime_data[,16]),
                         alpha = 1,
                         nfolds = 8,
                         nlambda = 20,
                         type.measure = "mse",
                         family = "gaussian",
                         standardize = TRUE)
model_lasso
```

```
##
## Call:  cv.glmnet(x = as.matrix(crime_data[, -16]), y = as.matrix(crime_data[,      16]), type.measur
##
## Measure: Mean-Squared Error
```

```
##
##     Lambda Index Measure    SE Nonzero
## min   8.84     8   62393 11474      11
## 1se  23.31     6   69651 13921       9
```

```r
#Plotting LASSO model
plot(model_lasso)
```




```r
#Finding Lambda minimum
model_lasso$lambda.min
```

```
## [1] 8.839527
```

```r
cbind(model_lasso$lambda, model_lasso$cvm, model_lasso$nzero)
```

```
##             [,1]      [,2] [,3]
## s0  263.09539664 148553.54    0
## s1  162.02682936 117144.60    1
## s2   99.78393301 102375.24    1
## s3   61.45175663  97832.60    4
## s4   37.84495439  80548.84    5
## s5   23.30674746  69650.63    9
## s6   14.35341873  63388.30   10
## s7    8.83952725  62392.77   11
```

```
## s8    5.44380704  64659.15   12
## s9    3.35255883  67693.93   12
## s10   2.06466736  70554.37   13
## s11   1.27152170  73867.35   14
## s12   0.78306436  76396.99   15
## s13   0.48224879  77983.53   15
## s14   0.29699205  78951.56   15
## s15   0.18290202  79407.81   15
## s16   0.11263988  79630.13   14
## s17   0.06936907  79710.17   15
## s18   0.04272082  79745.55   15
## s19   0.02630954  79782.02   15
```

```
coef(model_lasso, s = model_lasso$lambda.min)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -5.072255e+03
## M            7.184295e+01
## So           4.466407e+01
## Ed           1.253875e+02
## Po1          1.023402e+02
## Po2          .
## LF           .
## M.F          1.888147e+01
## Pop          .
## NW           6.315089e-01
## U1          -2.143645e+03
## U2           8.835503e+01
## Wealth       7.715072e-03
## Ineq         4.882548e+01
## Prob        -3.688177e+03
## Time         .
```

```
#Now using the above variables from the LASSO Regression results, I will
#Use the new factors.

model_lm <- lm( Crime~ M + So + Ed + Po1 + M.F + NW + U1 + U2 + Wealth + Ineq + Prob, data = crime_data)

summary(model_lm)
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + M.F + NW + U1 + U2 +
##     Wealth + Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -408.38  -96.14   -1.39  114.80  454.53
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.757e+03  1.313e+03  -5.147 1.03e-05 ***
```

```
## M             9.148e+01  3.893e+01   2.350  0.02454 *
## So            3.335e+01  1.237e+02   0.270  0.78905
## Ed            1.746e+02  5.589e+01   3.124  0.00357 **
## Po1           9.277e+01  2.019e+01   4.596 5.41e-05 ***
## M.F           2.189e+01  1.453e+01   1.506  0.14101
## NW            1.549e+00  5.559e+00   0.279  0.78209
## U1           -5.248e+03  3.600e+03  -1.458  0.15380
## U2            1.667e+02  7.853e+01   2.123  0.04089 *
## Wealth        7.626e-02  9.737e-02   0.783  0.43878
## Ineq          6.693e+01  2.022e+01   3.310  0.00217 **
## Prob         -3.854e+03  1.770e+03  -2.177  0.03627 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201.3 on 35 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7292
## F-statistic: 12.26 on 11 and 35 DF,  p-value: 5.334e-09
```

```r
#Looking at the factors based on their p-values, I can remove So, M.F, NW, U1, and Wealth.

new_lasso <- lm(Crime~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)

summary(new_lasso)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

After performing *Stepwise regression*, I performed *LASSO*and when using *LASSO*, it is essential to scale the data. Therefore, a line is added to the model to *standardize* it since the *LASSO* method does not force some coefficients to 0 in order to simplify the mode. I used an alpha value of 1 for variable selection. The model showed 11 significant factors but, with further investigation, the p-values were used to determine insignificant factors that could be eliminated to build a final regression model. The final model revealed factors identical to the *Stepwise regression* model.

Part 3 Elastic Net

```r
#From the notes in the question, it claims we will get a range of results by
#varying "alpha" from 1 (lasso) to 0 (ridge regression).
#So, for Elastic net, I will try different values and see which model gives the best result.

library(glmnet)

alpha <- seq(0, 1, 10 ) #range between 0 and 1
best <- list(a = NULL, mse = NULL)
for (i in 1:length(alpha)) {
  set.seed(42)
  elastic <- cv.glmnet(x = as.matrix(crime_data[,-16]), #Cross-validation
                       y = as.matrix(crime_data[,16]),
                       alpha = alpha[i],
                       nfolds = 10,
                       type.measure = "mse",
                       family = "gaussian")
  best$a <- c(best$a, alpha[i])
  best$mse <- c(best$mse, min(elastic$cvm))

 alpha[i] = elastic$glmnet.fit$dev.ratio[which(elastic$glmnet.fit$lambda
                                              == elastic$lambda.min)]
}
alpha[i]
```

```
## [1] 0.7493586
```

```r
best_alpha <- alpha[i]

index <- which(best$mse == min(best$mse))
best_mse1 <- best$mse[index]

cat("alpha:", best_alpha, "mse:", best_mse1) #results
```

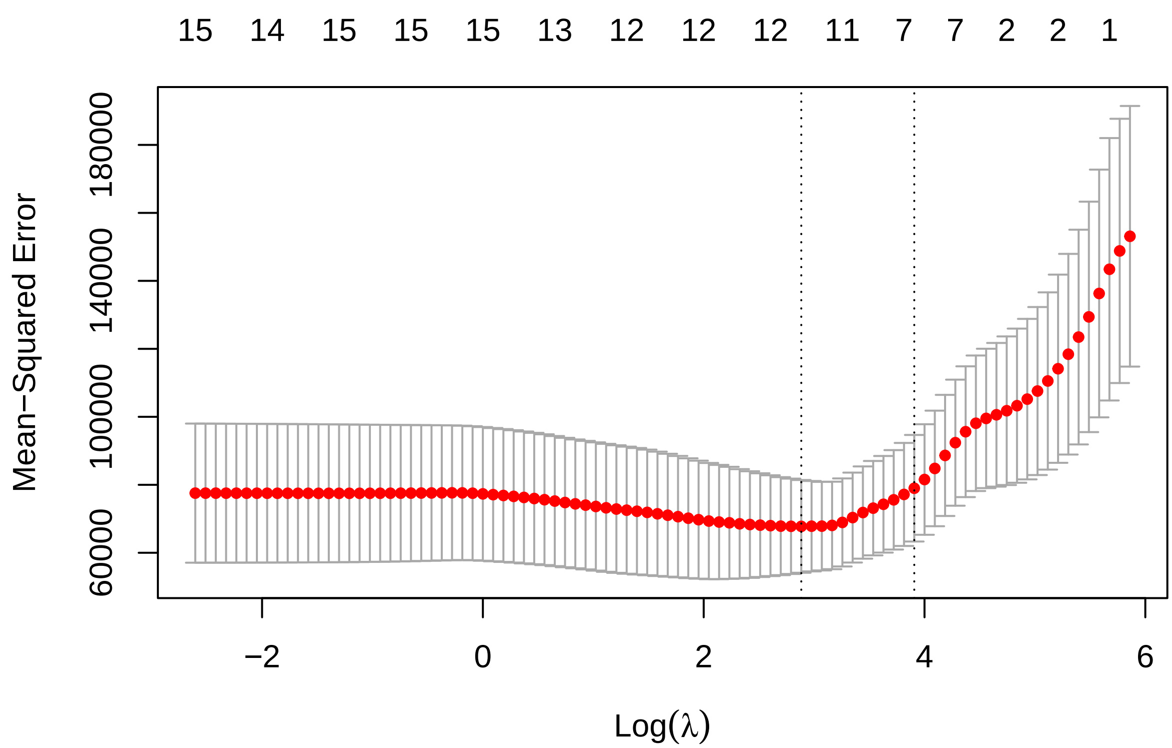
```
## alpha: 0.7493586 mse: 75272.52
```

```r
#Apply CV with best alpha to get lambda
new_elastic1 <- cv.glmnet(x = as.matrix(crime_data[,-16]),
                       y = as.matrix(crime_data[,16]),
                       alpha = best_alpha,
                       family = "gaussian")
summary(new_elastic1)
```

```
##            Length Class  Mode
## lambda     92     -none- numeric
## cvm        92     -none- numeric
## cvsd       92     -none- numeric
## cvup       92     -none- numeric
## cvlo       92     -none- numeric
## nzero      92     -none- numeric
```

```
## call        5      -none- call
## name        1      -none- character
## glmnet.fit 12      elnet  list
## lambda.min  1      -none- numeric
## lambda.1se  1      -none- numeric
## index       2      -none- numeric
```

```r
plot(new_elastic1)
```




```r
best_lambda <- new_elastic1$lambda.min
cat(best_lambda)
```

```
## 17.88522
```

```r
#Now we can fit model with the best alpha and lambda value
elastic_model <- glmnet(x = as.matrix(crime_data[,-16]),
                        y = as.matrix(crime_data[,16]),
                        alpha = best_alpha,
                        lambda = best_lambda,
                        family = "gaussian")
coef(elastic_model)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
```

```
## (Intercept) -4207.9348033
## M              60.4958342
## So             52.3283202
## Ed             89.0142710
## Po1            98.9587741
## Po2             2.2821932
## LF            125.9372409
## M.F            17.0780811
## Pop             .
## NW              0.8586093
## U1           -400.5915127
## U2             44.7235458
## Wealth          .
## Ineq           38.7178996
## Prob        -3523.4028465
## Time            .
```

```r
best_elastic <- lm(Crime~ M + So + Ed + Po1 + LF + M.F + NW + U2 + Ineq +
                     Prob, data = crime_data)
summary(best_elastic)
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U2 +
##     Ineq + Prob, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -410.55 -121.42    5.76  110.54  550.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5725.4999  1184.4847  -4.834 2.49e-05 ***
## M              90.4022    39.7548   2.274  0.02903 *
## So            122.7918   129.8958   0.945  0.35080
## Ed            168.7274    57.8799   2.915  0.00608 **
## Po1           112.2077    16.6064   6.757 6.86e-08 ***
## LF            622.6363  1227.0766   0.507  0.61496
## M.F            10.6943    14.6530   0.730  0.47021
## NW             -0.2803     5.7955  -0.048  0.96169
## U2             86.2961    49.4882   1.744  0.08973 .
## Ineq           57.4296    17.2560   3.328  0.00203 **
## Prob        -4360.5001  1769.9432  -2.464  0.01867 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 206.6 on 36 degrees of freedom
## Multiple R-squared:  0.7767, Adjusted R-squared:  0.7147
## F-statistic: 12.53 on 10 and 36 DF,  p-value: 5.374e-09
```

Here I used the Elastic net algorithm and set the alpha between 0 and 1 for variable selection. The model has more definitive selection than based on p values. The adjusted R^2 gives a 71% which is close to my Stepwise regression adjusted R^2 just more precise.