HW8

2022-10-17

#Step 1: Environment Setup This part is applicable to answer all three questions. First, we clear the environment, set the seed, load appropriate libraries, read data and create helper function to calculate R^2.

```
# Clear the environment
rm(list = ls())

# Comment in set.seed(33) to repeat results
set.seed(33)

# Load glmnet and DAAG lib
require(glmnet)
```

```
## Loading required package: glmnet

## Loading required package: Matrix

## Loaded glmnet 4.1-4
```

```
require(DAAG)
```

```
## Loading required package: DAAG
```

```
# Load crime data into a data frame
data_df <- read.table("uscrime.txt", header=TRUE)

# Scale the data and convert it to a matrix for LASSO and ELNET
scaled_data_df <- as.data.frame(scale(data_df[,c(1,3:15)]))
scaled_data_df <- cbind(data_df[,2],scaled_data_df,data_df[,16])
colnames(scaled_data_df)[1] <- "So"
colnames(scaled_data_df)[16] <- "Crime"
data_mx <- as.matrix(scaled_data_df)

# Helper function to calculate R^2 - will be used later
ComputeR2 <- function(yhat_df, data_df) {
  SSres <- sum((yhat_df - data_df$Crime)^2)
  SStot <- sum((data_df$Crime - mean(data_df$Crime))^2)
  R2 <- 1 - SSres/SStot
  return(R2)
}
```

#Stepwise Regression ##1: Identify Factors with Step(). I created a model with all components for stepwise regression factor selection. The step() function can be set to forward regression, backward regression, or both (stepwise) regression and selects a model by AIC in a Stepwise Algorithm. I ran step() on it with direction set to "both," which instructs step() to conduct "backward" and "forward" selection.

```r
# Create a linear regression model with all factors
model_all <- lm(Crime ~., data_df)
step(model_all, 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
```

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

##2: Retrain the Model Using the Factors Identified in Step(). I retrained the regression model after using step() to determine the factors to add in our model:

```
# Re-train model using "best" set of factors from step()
step_model <- lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob, data = data_df)
summary(step_model)
```

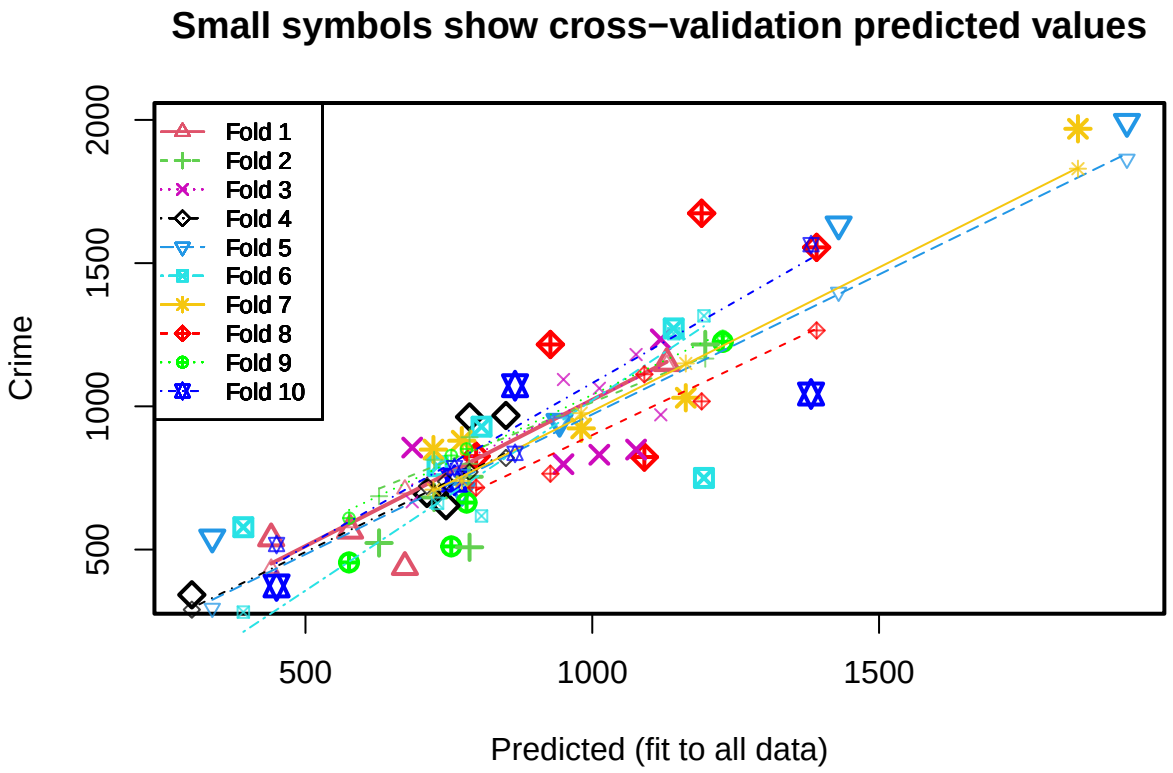
```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.70 -111.07    3.03  122.15  483.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6426.10    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
```

##3 Cross-Validate the Step-Model and Calculate R^2 Due to the small size of the dataset, I used cv.lm() to cross-validate the model and then utilized the cv prediction values to calculate R2 (using the ComputeR2 function defined in the Global Step above):

```
# Cross-validate the step_model
cv_step_model <- cv.lm(data = data_df, form.lm = step_model, m = 10)
```

```
## Warning in cv.lm(data = data_df, form.lm = step_model, m = 10):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
```

```
## fold 1
## Observations in test set: 4
##                    17        22        38          40
## Predicted   440.1808  673.3317 577.75909 1129.888624
## cvpred      434.7231  711.2440 593.54695 1147.548837
## Crime       539.0000  439.0000 566.00000 1151.000000
## CV residual 104.2769 -272.2440 -27.54695    3.451163
##
## Sum of squares = 85761.22    Mean square = 21440.3    n = 4
##
## fold 2
## Observations in test set: 5
##                     6        25         28        32        46
## Predicted    724.2856  628.2696 1197.00602 785.32166  786.0695
## cvpred       813.2059  686.2682 1166.63009 799.35994  892.1474
## Crime        682.0000  523.0000 1216.00000 754.00000  508.0000
## CV residual -131.2059 -163.2682   49.36991 -45.35994 -384.1474
##
## Sum of squares = 195935.6    Mean square = 39187.13    n = 5
##
## fold 3
## Observations in test set: 5
##                     5        9         15        37        47
## Predicted   1119.4533 686.1097  949.8039 1012.3317 1076.3622
## cvpred       970.3784 666.6828 1093.6907 1062.8204 1180.8007
## Crime       1234.0000 856.0000  798.0000  831.0000  849.0000
## CV residual  263.6216 189.3172 -295.6907 -231.8204 -331.8007
##
## Sum of squares = 356602.8    Mean square = 71320.55    n = 5
##
## fold 4
## Observations in test set: 5
##                     7       24        27        30        35
## Predicted   786.0570 849.5001 301.89278 711.81558 745.02008
## cvpred      770.6668 819.7856 290.11213 721.71531 737.31934
## Crime       963.0000 968.0000 342.00000 696.00000 653.00000
## CV residual 192.3332 148.2144  51.88787 -25.71531 -84.31934
##
## Sum of squares = 69422.95    Mean square = 13884.59    n = 5
##
## fold 5
## Observations in test set: 5
##                     2        10         16        26       42
## Predicted   1429.5290 772.69245 942.968516 1932.1846 337.5060
## cvpred      1398.6486 762.77888 944.588145 1863.1455 295.1243
## Crime       1635.0000 705.00000 946.000000 1993.0000 542.0000
## CV residual  236.3514 -57.77888   1.411855  129.8545 246.8757
##
## Sum of squares = 137012.2    Mean square = 27402.44    n = 5
##
## fold 6
## Observations in test set: 5
##                     1        3       18        19       36
## Predicted   730.2603 391.6707 806.9599 1194.7025 1142.001
```

```
## cvpred      661.4494 281.6416 617.2925 1315.8845 1253.443
## Crime       791.0000 578.0000 929.0000  750.0000 1272.000
## CV residual 129.5506 296.3584 311.7075 -565.8845   18.557
##
## Sum of squares = 522342.8    Mean square = 104468.6    n = 5
##
## fold 7
## Observations in test set: 5
##                    4       12        34       41        44
## Predicted   1846.750 723.1273 980.69542 772.4885 1163.0310
## cvpred      1829.594 709.0251 976.23303 739.5756 1149.7357
## Crime       1969.000 849.0000 923.00000 880.0000 1030.0000
## CV residual  139.406 139.9749 -53.23303 140.4244 -119.7357
##
## Sum of squares = 75916.43    Mean square = 15183.29    n = 5
##
## fold 8
## Observations in test set: 5
##                     8        11        23       39        43
## Predicted   1391.0999 1190.7017  927.0356 797.5843 1090.8352
## cvpred      1264.9034 1017.3621  764.8551 716.1954 1111.8991
## Crime       1555.0000 1674.0000 1216.0000 826.0000  823.0000
## CV residual  290.0966  656.6379  451.1449 109.8046 -288.8991
##
## Sum of squares = 814380.8    Mean square = 162876.2    n = 5
##
## fold 9
## Observations in test set: 4
##                    13        14         20        45
## Predicted    754.1956  780.8699 1227.55497  575.9466
## cvpred       828.3137  851.2144 1244.21765  609.7352
## Crime        511.0000  664.0000 1225.00000  455.0000
## CV residual -317.3137 -187.2144  -19.21765 -154.7352
##
## Sum of squares = 160049.5    Mean square = 40012.38    n = 4
##
## fold 10
## Observations in test set: 4
##                    21        29        31        33
## Predicted   759.79628 1381.4244  449.5679  865.3617
## cvpred      785.95176 1564.3767  517.6051  835.8392
## Crime       742.00000 1043.0000  373.0000 1072.0000
## CV residual -43.95176 -521.3767 -144.6051  236.1608
##
## Sum of squares = 350448    Mean square = 87611.99    n = 4
##
## Overall (Sum over all 4 folds)
##      ms
## 58890.9
```

```
# Calculate R^2 for the cv_step_model
step_yhat <- as.data.frame(cv_step_model$cvpred)
cv_step_model_R2 <- ComputeR2(step_yhat, data_df)
cv_step_model_R2
```

```
## [1] 0.5977472
```

According to the function, none of the additional factors would cause the AIC to fall. The model was equivalent to the initial stepwise regression model and no components were eliminated.

#LASSO

##1: Identify Factors using LASSO I found that the optimized lambda.min value was equivalent to 4.82 using cv.glmnet with an alpha = 1 and advised utilizing the following factors: So, M, Ed, Po1, M.F, Pop, NW, U1, U2, Wealth, Ineq, and Prop:

```
# Identify factors using LASSO
lasso_factors <- cv.glmnet(x = data_mx[,-16],
                           y = data_mx[,"Crime"],
                           alpha = 1,
                           nfolds = 5,
                           type.measure = "mse",
                           family = "gaussian")

# Display the lambda.min for lasso_factors
lasso_factors$lambda.min
```

```
## [1] 25.70468
```

```
# Display coefficients for lambda.min
lasso_coeff <- coef(lasso_factors, s = lasso_factors$lambda.min)
lasso_coeff
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 904.8624332
## So            0.6541023
## M            52.7646957
## Ed           25.7415683
## Po1         294.7913241
## Po2           .
## LF            .
## M.F          53.6009205
## Pop           .
## NW            2.6706728
## U1            .
## U2            1.9014702
## Wealth        .
## Ineq        111.0004689
## Prob        -58.9535406
## Time          .
```

##2: Retrain Model with LASSO Identified Factors I retrain the model based on the LASSO-recommended factors:

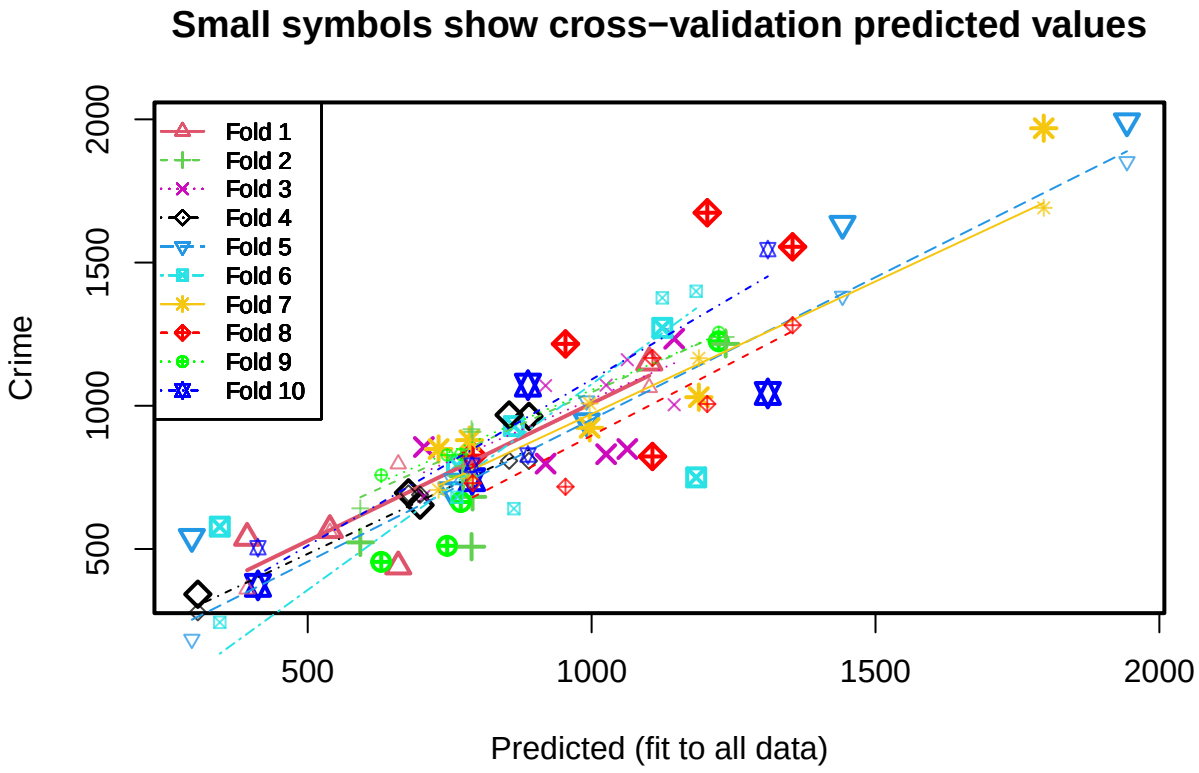
```
# Re-train model using lambda.min factors
lasso_model <- lm(formula = Crime ~ So + M + Ed + Po1 + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob
summary(lasso_model)
```

```
##
## Call:
## lm(formula = Crime ~ So + M + Ed + Po1 + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -434.18 -107.01   18.55  115.88  470.32
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.393e+03  1.413e+03  -4.524 7.05e-05 ***
## So           2.289e+01  1.253e+02   0.183  0.85621
## M            8.968e+01  3.927e+01   2.284  0.02876 *
## Ed           1.749e+02  5.627e+01   3.109  0.00378 **
## Po1          9.865e+01  2.187e+01   4.511 7.32e-05 ***
## M.F          1.660e+01  1.633e+01   1.017  0.31656
## Pop         -8.734e-01  1.199e+00  -0.729  0.47113
## NW           1.863e+00  5.613e+00   0.332  0.74195
## U1          -4.979e+03  3.643e+03  -1.367  0.18069
## U2           1.667e+02  7.906e+01   2.108  0.04245 *
## Wealth       8.633e-02  9.900e-02   0.872  0.38932
## Ineq         7.163e+01  2.135e+01   3.355  0.00196 **
## Prob        -4.079e+03  1.809e+03  -2.255  0.03065 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.6 on 34 degrees of freedom
## Multiple R-squared:  0.7971, Adjusted R-squared:  0.7255
## F-statistic: 11.13 on 12 and 34 DF,  p-value: 1.52e-08
```

##3: Cross-Validate the LASSO Model and Calculate R^2 Due to the small data set, I used cv.lm() to cross-validate the lasso model and calculated R2 using the cv prediction values (using the Compute R2 function defined in the Global Step above):

```
# Cross-validate the lasso_model
cv_lasso_model <- cv.lm(data = data_df, form.lm = lasso_model, m = 10)
```

```
## Warning in cv.lm(data = data_df, form.lm = lasso_model, m = 10):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 4
##                   17        22         38         40
## Predicted   393.2891  659.4696 539.032809 1101.57182
## cvpred      359.9602  796.8771 557.974318 1063.41303
## Crime       539.0000  439.0000 566.000000 1151.00000
## CV residual 179.0398 -357.8771   8.025682   87.58697
##
## Sum of squares = 167867.2    Mean square = 41966.8    n = 4
##
## fold 2
## Observations in test set: 5
##                     6        25         28        32        46
## Predicted    790.3269  592.6041 1236.13712 773.69706  788.5734
## cvpred       909.1160  641.8107 1240.12084 788.93474  918.0856
## Crime        682.0000  523.0000 1216.00000 754.00000  508.0000
## CV residual -227.1160 -118.8107  -24.12084 -34.93474 -410.0856
##
## Sum of squares = 235670.1    Mean square = 47134.02    n = 5
##
## fold 3
## Observations in test set: 5
##                     5       9        15        37        47
## Predicted   1145.4489 704.609  918.9447 1025.5824 1062.9107
## cvpred      1003.4644 684.036 1070.3131 1072.0197 1160.6902
```

```
## Crime       1234.0000 856.000  798.0000  831.0000  849.0000
## CV residual  230.5356 171.964 -272.3131 -241.0197 -311.6902
##
## Sum of squares = 312114    Mean square = 62422.8    n = 5
##
## fold 4
## Observations in test set: 5
##                    7       24        27         30        35
## Predicted   889.6438 854.7153 306.48625 677.447910 697.87598
## cvpred      807.3957 807.9612 279.15352 694.011773 689.42847
## Crime       963.0000 968.0000 342.00000 696.000000 653.00000
## CV residual 155.6043 160.0388  62.84648   1.988227 -36.42847
##
## Sum of squares = 55105.8    Mean square = 11021.16    n = 5
##
## fold 5
## Observations in test set: 5
##                     2        10         16        26       42
## Predicted   1441.9324 753.31682  991.26173 1942.8842 295.9088
## cvpred      1382.6012 746.27453 1017.12517 1852.5304 185.0087
## Crime       1635.0000 705.00000  946.00000 1993.0000 542.0000
## CV residual  252.3988 -41.27453  -71.12517  140.4696 356.9913
##
## Sum of squares = 217642    Mean square = 43528.41    n = 5
##
## fold 6
## Observations in test set: 5
##                    1        3       18        19        36
## Predicted   762.8805 345.1417 863.1375 1184.1776 1124.5564
## cvpred      680.7891 243.9700 640.8193 1399.6923 1376.6526
## Crime       791.0000 578.0000 929.0000  750.0000 1272.0000
## CV residual 110.2109 334.0300 288.1807 -649.6923 -104.6526
##
## Sum of squares = 639822.9    Mean square = 127964.6    n = 5
##
## fold 7
## Observations in test set: 5
##                     4       12        34       41       44
## Predicted   1796.3119 730.5284  996.6047 784.9130 1188.916
## cvpred      1691.0383 706.2872 1007.9765 746.0559 1166.042
## Crime       1969.0000 849.0000  923.0000 880.0000 1030.000
## CV residual  277.9617 142.7128  -84.9765 133.9441 -136.042
##
## Sum of squares = 141299.1    Mean square = 28259.82    n = 5
##
## fold 8
## Observations in test set: 5
##                    8        11        23        39        43
## Predicted   1353.963 1203.6845  953.9849 790.81637 1107.1750
## cvpred      1281.269 1006.1643  717.3450 729.99316 1166.8248
## Crime       1555.000 1674.0000 1216.0000 826.00000  823.0000
## CV residual  273.731  667.8357  498.6550  96.00684 -343.8248
##
## Sum of squares = 897022.7    Mean square = 179404.5    n = 5
```

```
##
## fold 9
## Observations in test set: 4
##                    13        14         20        45
## Predicted    745.5794  769.6995 1223.84363  629.4543
## cvpred       828.2562  830.4178 1256.19431  758.0607
## Crime        511.0000  664.0000 1225.00000  455.0000
## CV residual -317.2562 -166.4178  -31.19431 -303.0607
##
## Sum of squares = 221165.3    Mean square = 55291.32    n = 4
##
## fold 10
## Observations in test set: 4
##                    21        29        31        33
## Predicted   789.34290 1310.3439  412.3731  887.9003
## cvpred      792.65401 1545.2088  504.3217  828.0504
## Crime       742.00000 1043.0000  373.0000 1072.0000
## CV residual -50.65401 -502.2088 -131.3217  243.9496
##
## Sum of squares = 331536.3    Mean square = 82884.09    n = 4
##
## Overall (Sum over all 4 folds)
##       ms
## 68494.58
```

```
summary(cv_lasso_model)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob             Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
```

```
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
##    Predicted          cvpred            fold
##  Min.   : 295.9   Min.   : 185.0   Min.   : 1.000
##  1st Qu.: 717.6   1st Qu.: 700.1   1st Qu.: 3.000
##  Median : 854.7   Median : 828.1   Median : 5.000
##  Mean   : 905.1   Mean   : 907.1   Mean   : 5.426
##  3rd Qu.:1115.9   3rd Qu.:1116.4   3rd Qu.: 8.000
##  Max.   :1942.9   Max.   :1852.5   Max.   :10.000
```

```
# Calculate R^2 for the cv_step_model
lasso_yhat <- as.data.frame(cv_lasso_model$cvpred)
cv_lasso_model_R2 <- ComputeR2(lasso_yhat, data_df)
cv_lasso_model_R2
```

```
## [1] 0.5321495
```

#ELASTIC NET

##1: Variety of Alpha Values Similar to the lasso model, I created my elastic net model using the cv.glmnet function in the glmnet package. In an effort to change the alpha setting, I selected values of 0.25, 0.50, and 0.75 and conducted the following procedure: run cv.glmnet with a variety of alpha values, identify factors, retrain the model using the detected factors, cv, and calculate R2.

Using Alpha of 0.25

```
# Identify factors using Elastic Net and alpha of 0.25
elnet_factors <- cv.glmnet(x = data_mx[,-16],
                           y = data_mx[,"Crime"],
                           alpha = 0.25,
                           nfolds = 5,
                           type.measure = "mse",
                           family = "gaussian")

# Display the lambda.min for elnet_factors
elnet_factors$lambda.min
```

```
## [1] 33.6685
```

```
# Display the coefficients for lamdba.min
elnet_coeff <- coef(elnet_factors, s = elnet_factors$lambda.min)
elnet_coeff
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) 881.88683
## So           68.14493
## M            74.24322
## Ed           94.09007
## Po1         171.72034
## Po2         104.78977
## LF           15.67476
## M.F          59.87619
## Pop           .
```

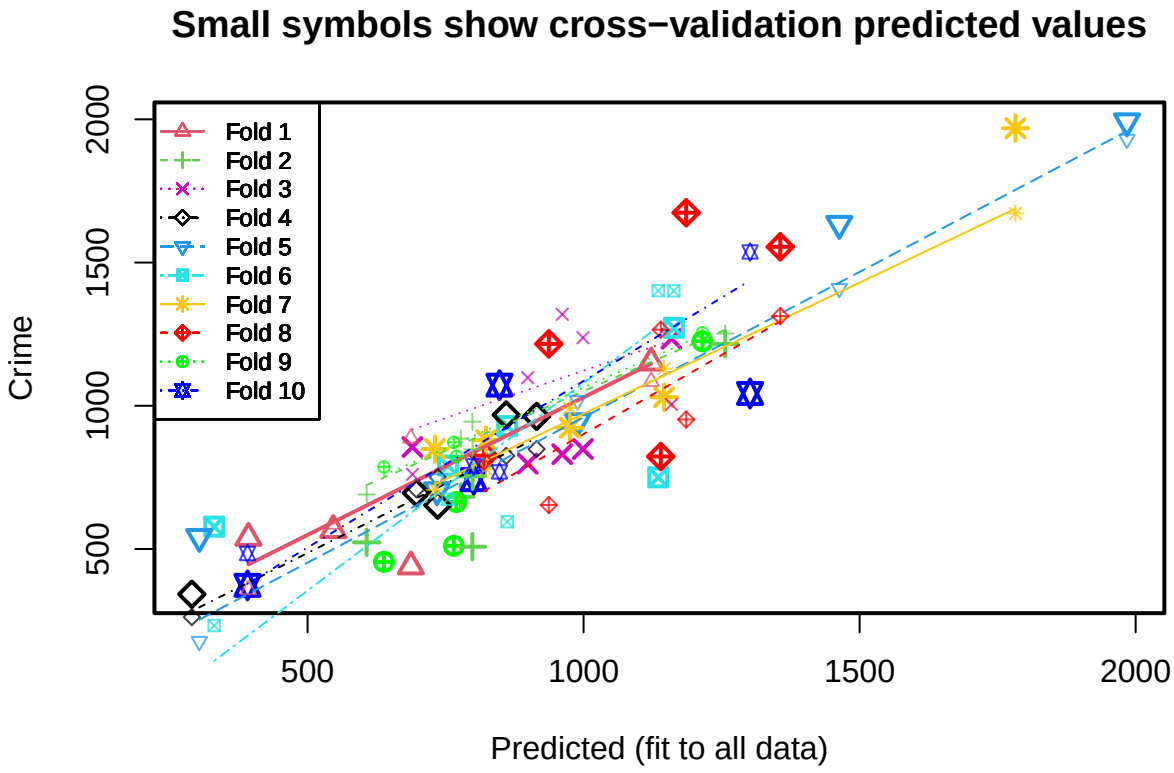
```
## NW           21.93618
## U1          -27.76295
## U2           61.03397
## Wealth       10.67713
## Ineq        135.39900
## Prob        -83.83310
## Time          .
```

```
# Re-train model using lambda.min factors
elnet_model <- lm(formula = Crime ~ So + M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + I
summary(elnet_model)
```

```
##
## Call:
## lm(formula = Crime ~ So + M + Ed + Po1 + Po2 + LF + M.F + Pop +
##     NW + U1 + U2 + Wealth + Ineq + Prob, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -385.20  -98.21    6.29  108.37  488.17
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.277e+03  1.495e+03  -4.200 0.000199 ***
## So           5.695e+00  1.457e+02   0.039 0.969060
## M            8.463e+01  4.070e+01   2.080 0.045656 *
## Ed           1.894e+02  6.131e+01   3.089 0.004134 **
## Po1          1.773e+02  9.995e+01   1.773 0.085664 .
## Po2         -8.932e+01  1.086e+02  -0.822 0.416972
## LF          -6.092e+02  1.448e+03  -0.421 0.676754
## M.F          1.913e+01  1.980e+01   0.966 0.341290
## Pop         -8.833e-01  1.237e+00  -0.714 0.480322
## NW           3.275e+00  6.117e+00   0.535 0.596110
## U1          -5.550e+03  4.121e+03  -1.347 0.187530
## U2           1.636e+02  8.090e+01   2.023 0.051546 .
## Wealth       9.042e-02  1.017e-01   0.889 0.380848
## Ineq         7.091e+01  2.244e+01   3.160 0.003434 **
## Prob        -4.232e+03  1.853e+03  -2.284 0.029132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 206.6 on 32 degrees of freedom
## Multiple R-squared:  0.8016, Adjusted R-squared:  0.7148
## F-statistic: 9.234 on 14 and 32 DF,  p-value: 1.249e-07
```

```
# Cross-validate the elnet_model
cv_elnet_model <- cv.lm(data = data_df, form.lm = elnet_model, m = 10)
```

```
## Warning in cv.lm(data = data_df, form.lm = elnet_model, m = 10):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 4
##                   17        22        38         40
## Predicted   392.9025  687.0811 546.67624 1122.36362
## cvpred      357.8009  887.1465 586.11382 1085.04087
## Crime       539.0000  439.0000 566.00000 1151.00000
## CV residual 181.1991 -448.1465 -20.11382   65.95913
##
## Sum of squares = 238423.6    Mean square = 59605.9    n = 4
##
## fold 2
## Observations in test set: 5
##                     6        25         28        32        46
## Predicted    777.6851  607.0108 1256.49119 798.91966  798.5025
## cvpred       885.5354  690.1308 1252.22154 847.94139  944.6172
## Crime        682.0000  523.0000 1216.00000 754.00000  508.0000
## CV residual -203.5354 -167.1308  -36.22154 -93.94139 -436.6172
##
## Sum of squares = 270130.9    Mean square = 54026.19    n = 5
##
## fold 3
## Observations in test set: 5
##                     5         9        15        37        47
## Predicted   1159.1829 689.90317  898.7378  961.3220  998.8340
## cvpred      1004.6275 760.53539 1098.0976 1319.3179 1237.6747
```

```
## Crime       1234.0000 856.00000  798.0000  831.0000  849.0000
## CV residual  229.3725  95.46461 -300.0976 -488.3179 -388.6747
##
## Sum of squares = 541306.2    Mean square = 108261.2    n = 5
##
## fold 4
## Observations in test set: 5
##                    7       24        27        30        35
## Predicted   914.8182 859.7906 290.35640 696.97680 735.80364
## cvpred      848.8177 825.0941 262.50468 708.78297 750.63741
## Crime       963.0000 968.0000 342.00000 696.00000 653.00000
## CV residual 114.1823 142.9059  79.49532 -12.78297 -97.63741
##
## Sum of squares = 49475.66    Mean square = 9895.13    n = 5
##
## fold 5
## Observations in test set: 5
##                     2        10         16         26       42
## Predicted   1463.2036 734.46312  989.22179 1984.26789 303.8838
## cvpred      1409.6686 745.23111 1018.07041 1930.66152 177.4878
## Crime       1635.0000 705.00000  946.00000 1993.00000 542.0000
## CV residual  225.3314 -40.23111  -72.07041   62.33848 364.5122
##
## Sum of squares = 194342.2    Mean square = 38868.43    n = 5
##
## fold 6
## Observations in test set: 5
##                    1        3       18        19        36
## Predicted   754.2563 330.8650 861.6102 1135.2036 1163.4770
## cvpred      673.5092 232.8051 595.1567 1401.9936 1401.3894
## Crime       791.0000 578.0000 929.0000  750.0000 1272.0000
## CV residual 117.4908 345.1949 333.8433 -651.9936 -129.3894
##
## Sum of squares = 686252.2    Mean square = 137250.4    n = 5
##
## fold 7
## Observations in test set: 5
##                     4       12        34       41         44
## Predicted   1782.3696 730.8214 975.05543 823.2075 1145.74933
## cvpred      1671.5352 706.7914 992.16237 779.4351 1129.49725
## Crime       1969.0000 849.0000 923.00000 880.0000 1030.00000
## CV residual  297.4648 142.2086 -69.16237 100.5649  -99.49725
##
## Sum of squares = 133505    Mean square = 26701    n = 5
##
## fold 8
## Observations in test set: 5
##                     8        11        23        39        43
## Predicted   1356.3101 1185.8318  936.9854 819.70688 1139.8894
## cvpred      1313.0489  952.0546  653.5553 805.51312 1265.4617
## Crime       1555.0000 1674.0000 1216.0000 826.00000  823.0000
## CV residual  241.9511  721.9454  562.4447  20.48688 -442.4617
##
## Sum of squares = 1092282    Mean square = 218456.3    n = 5
```

```
##
## fold 9
## Observations in test set: 4
##                    13        14         20        45
## Predicted    764.8018  769.4923 1215.26479  638.2623
## cvpred       872.1858  823.2202 1255.01773  786.3503
## Crime        511.0000  664.0000 1225.00000  455.0000
## CV residual -361.1858 -159.2202  -30.01773 -331.3503
##
## Sum of squares = 266500.4    Mean square = 66625.09    n = 4
##
## fold 10
## Observations in test set: 4
##                    21        29        31        33
## Predicted   801.03044 1301.5449  391.1906  847.6753
## cvpred      790.07082 1536.9698  484.2290  769.0291
## Crime       742.00000 1043.0000  373.0000 1072.0000
## CV residual -48.07082 -493.9698 -111.2290  302.9709
##
## Sum of squares = 350480.2    Mean square = 87620.06    n = 4
##
## Overall (Sum over all 4 folds)
##    ms
## 81334
```

```
summary(cv_elnet_model)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob              Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
```

```
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
##    Predicted          cvpred            fold
##  Min.   : 290.4   Min.   : 177.5   Min.   : 1.000
##  1st Qu.: 732.6   1st Qu.: 727.0   1st Qu.: 3.000
##  Median : 847.7   Median : 848.8   Median : 5.000
##  Mean   : 905.1   Mean   : 926.1   Mean   : 5.426
##  3rd Qu.:1137.5   3rd Qu.:1183.6   3rd Qu.: 8.000
##  Max.   :1984.3   Max.   :1930.7   Max.   :10.000
```

```
# Calculate R^2 for the cv_elnet_model
elnet_yhat <- as.data.frame(cv_elnet_model$cvpred)
cv_elnet_model_R2 <- ComputeR2(elnet_yhat, data_df)
cv_elnet_model_R2
```

```
## [1] 0.4444502
```

Using Alpha of 0.5

```
# Identify factors using Elastic Net and alpha of 0.50
elnet_factors <- cv.glmnet(x = data_mx[,-16],
                           y = data_mx[,"Crime"],
                           alpha = 0.50,
                           nfolds = 5,
                           type.measure = "mse",
                           family = "gaussian")

# Display the lambda.min for elnet_factors
elnet_factors$lambda.min
```

```
## [1] 24.42361
```

```
# Display the coefficients for lamdba.min
elnet_coeff <- coef(elnet_factors, s = elnet_factors$lambda.min)
elnet_coeff
```

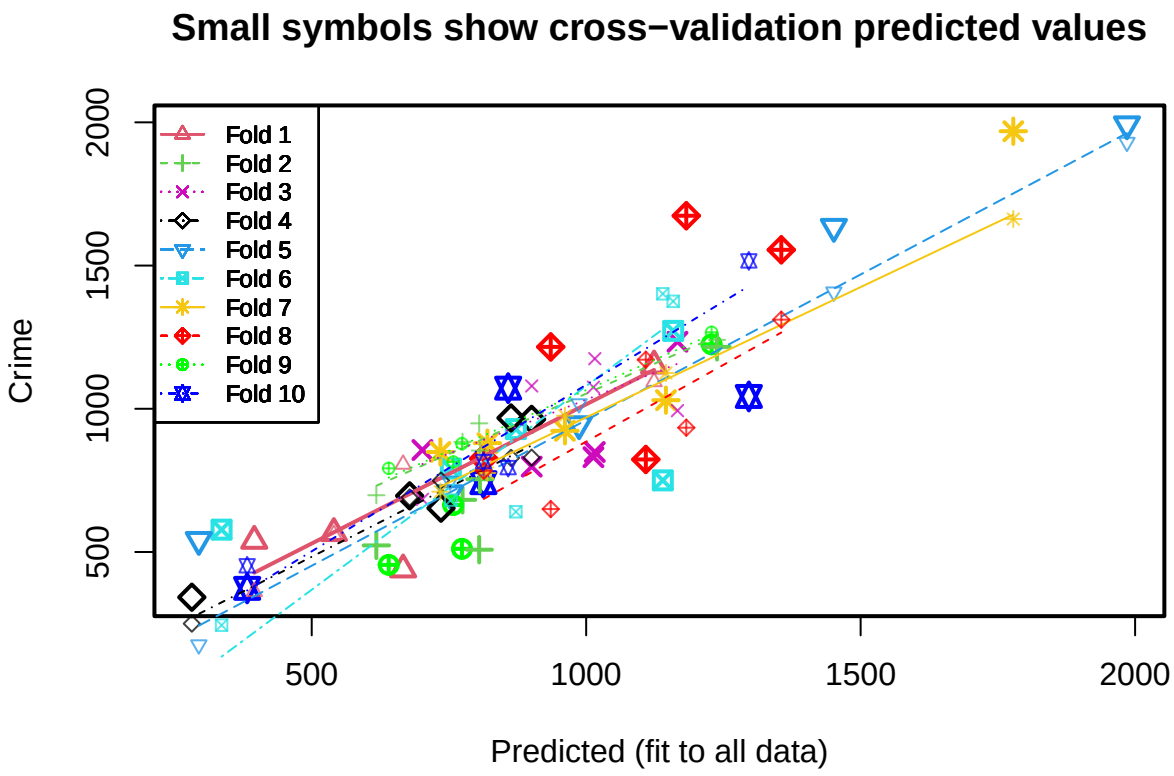
```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                     s1
## (Intercept) 885.24888
## So           58.26891
## M            73.99639
## Ed           92.41091
## Po1         210.17964
## Po2          82.30348
## LF           11.07080
## M.F          53.99072
## Pop           .
## NW           12.89806
## U1          -13.43028
## U2           45.42707
## Wealth        .
## Ineq        142.29202
## Prob        -81.00777
## Time          .
```

```r
# Re-train model using lambda.min factors
elnet_model <- lm(formula = Crime ~ So + M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth + Ineq +
summary(elnet_model)
```

```
##
## Call:
## lm(formula = Crime ~ So + M + Ed + Po1 + Po2 + M.F + Pop + NW +
##     U1 + U2 + Wealth + Ineq + Prob, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -389.63  -94.25    7.83  109.20  491.62
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.169e+03  1.454e+03  -4.243 0.000168 ***
## So           3.440e+01  1.271e+02   0.271 0.788398
## M            8.743e+01  3.964e+01   2.205 0.034514 *
## Ed           1.809e+02  5.721e+01   3.163 0.003346 **
## Po1          1.688e+02  9.667e+01   1.746 0.090115 .
## Po2         -7.692e+01  1.032e+02  -0.745 0.461484
## M.F          1.474e+01  1.663e+01   0.887 0.381622
## Pop         -9.510e-01  1.211e+00  -0.785 0.437837
## NW           2.422e+00  5.699e+00   0.425 0.673604
## U1          -4.805e+03  3.674e+03  -1.308 0.200017
## U2           1.622e+02  7.982e+01   2.032 0.050269 .
## Wealth       8.501e-02  9.967e-02   0.853 0.399833
## Ineq         6.912e+01  2.175e+01   3.177 0.003219 **
## Prob        -4.185e+03  1.826e+03  -2.292 0.028430 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204 on 33 degrees of freedom
## Multiple R-squared:  0.8005, Adjusted R-squared:  0.7219
## F-statistic: 10.19 on 13 and 33 DF,  p-value: 4.088e-08
```

```r
# Cross-validate the elnet_model
cv_elnet_model <- cv.lm(data = data_df, form.lm = elnet_model, m = 10)
```

```
## Warning in cv.lm(data = data_df, form.lm = elnet_model, m = 10):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 4
##                   17        22        38         40
## Predicted   395.3425  666.4836 540.40575 1123.69869
## cvpred      361.7375  806.3040 561.20577 1094.38323
## Crime       539.0000  439.0000 566.00000 1151.00000
## CV residual 177.2625 -367.3040   4.79423   56.61677
##
## Sum of squares = 169562.7    Mean square = 42390.67    n = 4
##
## fold 2
## Observations in test set: 5
##                     6        25         28        32        46
## Predicted    774.7670  617.7197 1238.39868  806.4278  805.0895
## cvpred       884.4192  698.1262 1240.55191  854.1236  949.0824
## Crime        682.0000  523.0000 1216.00000  754.0000  508.0000
## CV residual -202.4192 -175.1262  -24.55191 -100.1236 -441.0824
##
## Sum of squares = 276824    Mean square = 55364.79    n = 5
##
## fold 3
## Observations in test set: 5
##                    5        9        15        37        47
## Predicted   1166.349 701.6709  900.8028 1013.2388 1015.7527
## cvpred       993.418 683.7331 1079.6344 1075.3384 1174.8298
```

```
## Crime       1234.000 856.0000  798.0000  831.0000  849.0000
## CV residual  240.582 172.2669 -281.6344 -244.3384 -325.8298
##
## Sum of squares = 332739.8    Mean square = 66547.97    n = 5
##
## fold 4
## Observations in test set: 5
##                    7       24        27         30        35
## Predicted   900.5985 863.0837 281.63606 678.431771 735.51892
## cvpred      829.2449 828.4376 249.89166 690.081037 749.18333
## Crime       963.0000 968.0000 342.00000 696.000000 653.00000
## CV residual 133.7551 139.5624  92.10834   5.918963 -96.18333
##
## Sum of squares = 55138.3    Mean square = 11027.66    n = 5
##
## fold 5
## Observations in test set: 5
##                     2       10         16         26       42
## Predicted   1451.2197 751.9105  987.09261 1985.16584 294.3338
## cvpred      1407.9602 747.6052 1017.59529 1930.48193 175.9606
## Crime       1635.0000 705.0000  946.00000 1993.00000 542.0000
## CV residual  227.0398 -42.6052  -71.59529   62.51807 366.0394
##
## Sum of squares = 196381.5    Mean square = 39276.31    n = 5
##
## fold 6
## Observations in test set: 5
##                    1        3       18        19        36
## Predicted   753.4901 335.8127 871.9225 1139.6306 1158.5224
## cvpred      681.3245 244.4505 640.5086 1401.3299 1375.0966
## Crime       791.0000 578.0000 929.0000  750.0000 1272.0000
## CV residual 109.6755 333.5495 288.4914 -651.3299 -103.0966
##
## Sum of squares = 641370.8    Mean square = 128274.2    n = 5
##
## fold 7
## Observations in test set: 5
##                     4       12       34       41         44
## Predicted   1778.1171 734.4852 961.3284 820.0758 1145.32382
## cvpred      1662.5514 710.4871 975.6429 775.7140 1124.94139
## Crime       1969.0000 849.0000 923.0000 880.0000 1030.00000
## CV residual  306.4486 138.5129 -52.6429 104.2860  -94.94139
##
## Sum of squares = 135757.3    Mean square = 27151.45    n = 5
##
## fold 8
## Observations in test set: 5
##                     8        11        23        39        43
## Predicted   1355.4615 1182.3794  935.4453 813.45679 1108.6707
## cvpred      1311.1132  934.0392  649.8461 783.81684 1171.3516
## Crime       1555.0000 1674.0000 1216.0000 826.00000  823.0000
## CV residual  243.8868  739.9608  566.1539  42.18316 -348.3516
##
## Sum of squares = 1050681    Mean square = 210136.3    n = 5
```

```
##
## fold 9
## Observations in test set: 4
##                    13        14         20        45
## Predicted    773.7357  757.7743 1228.47989  640.3076
## cvpred       878.3367  814.2980 1267.09957  792.0317
## Crime        511.0000  664.0000 1225.00000  455.0000
## CV residual -367.3367 -150.2980  -42.09957 -337.0317
##
## Sum of squares = 272888.4    Mean square = 68222.11    n = 4
##
## fold 10
## Observations in test set: 4
##                    21        29        31        33
## Predicted   812.84060 1296.3383 382.40212  857.8604
## cvpred      815.46439 1516.2674 451.78758  794.5837
## Crime       742.00000 1043.0000 373.00000 1072.0000
## CV residual -73.46439 -473.2674 -78.78758  277.4163
##
## Sum of squares = 312546.3    Mean square = 78136.57    n = 4
##
## Overall (Sum over all 4 folds)
##       ms
## 73274.27
```

```
summary(cv_elnet_model)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob             Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
```

```
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
##    Predicted          cvpred            fold
##  Min.   : 281.6   Min.   : 176.0   Min.   : 1.000
##  1st Qu.: 735.0   1st Qu.: 704.3   1st Qu.: 3.000
##  Median : 857.9   Median : 829.2   Median : 5.000
##  Mean   : 905.1   Mean   : 911.8   Mean   : 5.426
##  3rd Qu.:1131.7   3rd Qu.:1109.7   3rd Qu.: 8.000
##  Max.   :1985.2   Max.   :1930.5   Max.   :10.000
```

```
# Calculate R^2 for the cv_elnet_model
elnet_yhat <- as.data.frame(cv_elnet_model$cvpred)
cv_elnet_model_R2 <- ComputeR2(elnet_yhat, data_df)
cv_elnet_model_R2
```

```
## [1] 0.499502
```

Using Alpha of 0.75

```
# Identify factors using Elastic Net and alpha of 0.75
elnet_factors <- cv.glmnet(x = data_mx[,-16],
                           y = data_mx[,"Crime"],
                           alpha = 0.75,
                           nfolds = 5,
                           type.measure = "mse",
                           family = "gaussian")

# Display the lambda.min for elnet_factors
elnet_factors$lambda.min
```

```
## [1] 28.45397
```

```
# Display the coefficients for lamdba.min
elnet_coeff <- coef(elnet_factors, s = elnet_factors$lambda.min)
elnet_coeff
```

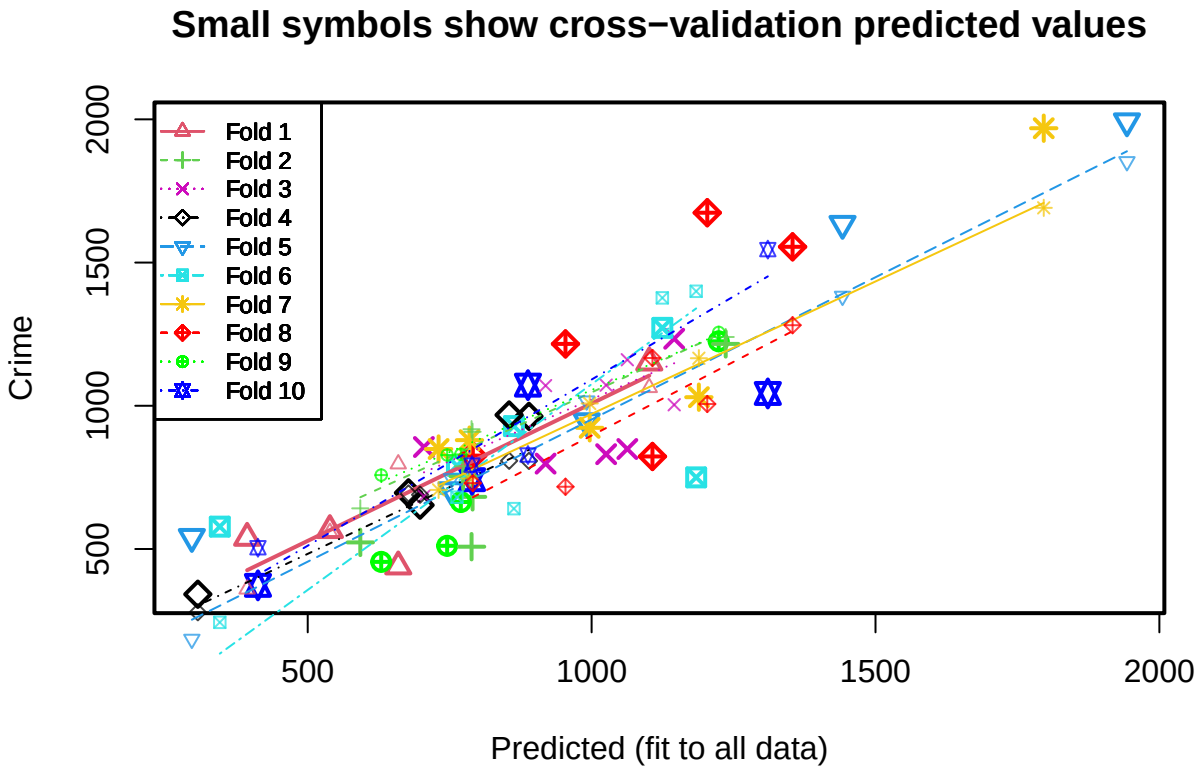
```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                     s1
## (Intercept) 897.93551
## So           21.00193
## M            57.39735
## Ed           49.03170
## Po1         252.91656
## Po2          35.31379
## LF            .
## M.F          53.83819
## Pop           .
## NW           10.43398
## U1            .
## U2           12.04028
## Wealth        .
## Ineq        116.43674
## Prob        -68.19152
## Time          .
```

```r
# Re-train model using lambda.min factors
elnet_model <- lm(formula = Crime ~ So + M + Ed + Po1 + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob
summary(elnet_model)
```

```
##
## Call:
## lm(formula = Crime ~ So + M + Ed + Po1 + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob, data = data_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -434.18 -107.01   18.55  115.88  470.32
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.393e+03  1.413e+03  -4.524 7.05e-05 ***
## So           2.289e+01  1.253e+02   0.183  0.85621
## M            8.968e+01  3.927e+01   2.284  0.02876 *
## Ed           1.749e+02  5.627e+01   3.109  0.00378 **
## Po1          9.865e+01  2.187e+01   4.511 7.32e-05 ***
## M.F          1.660e+01  1.633e+01   1.017  0.31656
## Pop         -8.734e-01  1.199e+00  -0.729  0.47113
## NW           1.863e+00  5.613e+00   0.332  0.74195
## U1          -4.979e+03  3.643e+03  -1.367  0.18069
## U2           1.667e+02  7.906e+01   2.108  0.04245 *
## Wealth       8.633e-02  9.900e-02   0.872  0.38932
## Ineq         7.163e+01  2.135e+01   3.355  0.00196 **
## Prob        -4.079e+03  1.809e+03  -2.255  0.03065 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.6 on 34 degrees of freedom
## Multiple R-squared:  0.7971, Adjusted R-squared:  0.7255
## F-statistic: 11.13 on 12 and 34 DF,  p-value: 1.52e-08
```

```r
# Cross-validate the elnet_model
cv_elnet_model <- cv.lm(data = data_df, form.lm = elnet_model, m = 10)
```

```
## Warning in cv.lm(data = data_df, form.lm = elnet_model, m = 10):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 4
##                   17        22         38         40
## Predicted   393.2891  659.4696 539.032809 1101.57182
## cvpred      359.9602  796.8771 557.974318 1063.41303
## Crime       539.0000  439.0000 566.000000 1151.00000
## CV residual 179.0398 -357.8771   8.025682   87.58697
##
## Sum of squares = 167867.2    Mean square = 41966.8    n = 4
##
## fold 2
## Observations in test set: 5
##                     6        25         28        32        46
## Predicted    790.3269  592.6041 1236.13712 773.69706  788.5734
## cvpred       909.1160  641.8107 1240.12084 788.93474  918.0856
## Crime        682.0000  523.0000 1216.00000 754.00000  508.0000
## CV residual -227.1160 -118.8107  -24.12084 -34.93474 -410.0856
##
## Sum of squares = 235670.1    Mean square = 47134.02    n = 5
##
## fold 3
## Observations in test set: 5
##                     5       9        15        37        47
## Predicted   1145.4489 704.609  918.9447 1025.5824 1062.9107
## cvpred      1003.4644 684.036 1070.3131 1072.0197 1160.6902
```

```
## Crime       1234.0000 856.000  798.0000  831.0000  849.0000
## CV residual  230.5356 171.964 -272.3131 -241.0197 -311.6902
##
## Sum of squares = 312114    Mean square = 62422.8    n = 5
##
## fold 4
## Observations in test set: 5
##                    7       24        27         30        35
## Predicted   889.6438 854.7153 306.48625 677.447910 697.87598
## cvpred      807.3957 807.9612 279.15352 694.011773 689.42847
## Crime       963.0000 968.0000 342.00000 696.000000 653.00000
## CV residual 155.6043 160.0388  62.84648   1.988227 -36.42847
##
## Sum of squares = 55105.8    Mean square = 11021.16    n = 5
##
## fold 5
## Observations in test set: 5
##                     2        10         16        26       42
## Predicted   1441.9324 753.31682  991.26173 1942.8842 295.9088
## cvpred      1382.6012 746.27453 1017.12517 1852.5304 185.0087
## Crime       1635.0000 705.00000  946.00000 1993.0000 542.0000
## CV residual  252.3988 -41.27453  -71.12517  140.4696 356.9913
##
## Sum of squares = 217642    Mean square = 43528.41    n = 5
##
## fold 6
## Observations in test set: 5
##                    1        3       18        19        36
## Predicted   762.8805 345.1417 863.1375 1184.1776 1124.5564
## cvpred      680.7891 243.9700 640.8193 1399.6923 1376.6526
## Crime       791.0000 578.0000 929.0000  750.0000 1272.0000
## CV residual 110.2109 334.0300 288.1807 -649.6923 -104.6526
##
## Sum of squares = 639822.9    Mean square = 127964.6    n = 5
##
## fold 7
## Observations in test set: 5
##                     4       12        34       41       44
## Predicted   1796.3119 730.5284  996.6047 784.9130 1188.916
## cvpred      1691.0383 706.2872 1007.9765 746.0559 1166.042
## Crime       1969.0000 849.0000  923.0000 880.0000 1030.000
## CV residual  277.9617 142.7128  -84.9765 133.9441 -136.042
##
## Sum of squares = 141299.1    Mean square = 28259.82    n = 5
##
## fold 8
## Observations in test set: 5
##                     8        11        23        39        43
## Predicted   1353.963 1203.6845  953.9849 790.81637 1107.1750
## cvpred      1281.269 1006.1643  717.3450 729.99316 1166.8248
## Crime       1555.000 1674.0000 1216.0000 826.00000  823.0000
## CV residual  273.731  667.8357  498.6550  96.00684 -343.8248
##
## Sum of squares = 897022.7    Mean square = 179404.5    n = 5
```

```
##
## fold 9
## Observations in test set: 4
##                    13        14         20        45
## Predicted    745.5794  769.6995 1223.84363  629.4543
## cvpred       828.2562  830.4178 1256.19431  758.0607
## Crime        511.0000  664.0000 1225.00000  455.0000
## CV residual -317.2562 -166.4178  -31.19431 -303.0607
##
## Sum of squares = 221165.3    Mean square = 55291.32    n = 4
##
## fold 10
## Observations in test set: 4
##                    21        29        31        33
## Predicted   789.34290 1310.3439  412.3731  887.9003
## cvpred      792.65401 1545.2088  504.3217  828.0504
## Crime       742.00000 1043.0000  373.0000 1072.0000
## CV residual -50.65401 -502.2088 -131.3217  243.9496
##
## Sum of squares = 331536.3    Mean square = 82884.09    n = 4
##
## Overall (Sum over all 4 folds)
##       ms
## 68494.58
```

```
summary(cv_elnet_model)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1               U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob             Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
```

```
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
##    Predicted          cvpred            fold
##  Min.   : 295.9   Min.   : 185.0   Min.   : 1.000
##  1st Qu.: 717.6   1st Qu.: 700.1   1st Qu.: 3.000
##  Median : 854.7   Median : 828.1   Median : 5.000
##  Mean   : 905.1   Mean   : 907.1   Mean   : 5.426
##  3rd Qu.:1115.9   3rd Qu.:1116.4   3rd Qu.: 8.000
##  Max.   :1942.9   Max.   :1852.5   Max.   :10.000
```

```
# Calculate R^2 for the cv_elnet_model
elnet_yhat <- as.data.frame(cv_elnet_model$cvpred)
cv_elnet_model_R2 <- ComputeR2(elnet_yhat, data_df)
cv_elnet_model_R2
```

```
## [1] 0.5321495
```

In conclusion, I discovered that the LASSO and Elastic Net methods indicated a large number of parameters that would probably result in over fitting. This model would be made more straightforward by eliminating some of the less significant components if I were creating it for a production system.