```
> set.seed(1)
>
> #install.packages("stats")
> #install.packages("DAAG")
> library(stats)
> library(DAAG)
>
> input = data.frame(read.table("uscrime.txt", header = TRUE)) #read in data
> mydata = input[c(16, 1:15)] #reorder so that crime is the first column (for formula)
>
> # Plot predictors vs. response
> predictors = mydata[-1]
> headers = list(
+   "M",
+   "So",
+   "Ed",
+   "Po1",
+   "Po2",
+   "LF",
+   "M.F",
+   "Pop",
+   "NW",
+   "U1",
+   "U2",
+   "Wealth",
+   "Ineq",
+   "Prob",
+   "Time"
+ )
> par(mfrow = c(4, 4))
> for (i in 1:15) {
+   plot(predictors[, i], mydata$Crime, xlab = headers[i])
+ }
>
>
> point = data.frame(
+   M = 14.0,
+   So = 0,
+   Ed = 10.0,
+   Po1 = 12.0,
+   Po2 = 15.5,
+   LF = 0.640,
+   M.F = 94.0,
+   Pop = 150,
+   NW = 1.1,
+   U1 = 0.120,
+   U2 = 3.6,
+   Wealth = 3200,
+   Ineq = 20.1,
+   Prob = 0.04,
+   Time = 39.0
+ )
>
>
> f1 = formula(mydata)
> model1 = lm(f1, mydata)
> summary(model1)

Call:
lm(formula = f1, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-395.7  -98.1   -6.7  113.0  512.7

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.98e+03   1.63e+03   -3.68  0.00089 ***
M            8.78e+01   4.17e+01    2.11  0.04344 *
So          -3.80e+00   1.49e+02   -0.03  0.97977
Ed           1.88e+02   6.21e+01    3.03  0.00486 **
Po1          1.93e+02   1.06e+02    1.82  0.07889 .
Po2         -1.09e+02   1.17e+02   -0.93  0.35883
LF          -6.64e+02   1.47e+03   -0.45  0.65465
M.F          1.74e+01   2.04e+01    0.86  0.39900
Pop         -7.33e-01   1.29e+00   -0.57  0.57385
NW           4.20e+00   6.48e+00    0.65  0.52128
U1          -5.83e+03   4.21e+03   -1.38  0.17624
U2           1.68e+02   8.23e+01    2.04  0.05016 .
Wealth       9.62e-02   1.04e-01    0.93  0.36075
Ineq         7.07e+01   2.27e+01    3.11  0.00398 **
Prob        -4.86e+03   2.27e+03   -2.14  0.04063 *
Time        -3.48e+00   7.17e+00   -0.49  0.63071
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209 on 31 degrees of freedom
Multiple R-squared:  0.803, Adjusted R-squared:  0.708
F-statistic: 8.43 on 15 and 31 DF,  p-value: 3.54e-07

> coef1 = model1$coefficients
>
> par(mfrow = c(2, 2))
> plot(model1)
>
> crime_prediction = predict.lm(model1, point)
> crime_prediction
  1
155
> #that answer of 155 is really low, we are probably overfit, so let's look at the p-values of each point.
>
>
> Pvalues = summary(model1)$coefficients[, 4]
> coef = model1$coefficients
>
>
> # Eliminate those predictors with a p-value > 0.08.  I know  0.05 is usually the rule,
> # but the U2 factor (unemployment rate of urban males 35-39) and Po1
> # should be left in as it was considered important by the summary() function.
> mydata_fit = mydata[1]
> n = 2
> for (i in 2:16) {
+   if (Pvalues[i] < 0.08) {
+     mydata_fit[n] = mydata[i]
+     n = n + 1
+   }
+ }
>
> f2 = formula(mydata_fit)
> model2 = lm(f2, mydata_fit)
> summary(model2)

Call:
lm(formula = f2, data = mydata_fit)

Residuals:
   Min     1Q Median     3Q    Max
-470.7  -78.4  -19.7  133.1  556.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
M              105.0       33.3    3.15   0.0031 **
Ed             196.5       44.8    4.39  8.1e-05 ***
Po1            115.0       13.8    8.36  2.6e-10 ***
U2              89.4       40.9    2.18   0.0348 *
Ineq            67.7       13.9    4.85  1.9e-05 ***
Prob         -3801.8     1528.1   -2.49   0.0171 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 201 on 40 degrees of freedom
Multiple R-squared:  0.766, Adjusted R-squared:  0.731
F-statistic: 21.8 on 6 and 40 DF,  p-value: 3.42e-11

> plot(model2)
>
>
> crime_prediction_adj = predict.lm(model2, point)
>
> crime_prediction_adj
   1
1304
>
> #Try cross-validation as well:
> par(mfrow = c(1, 1))
> c1 = cv.lm(mydata, model1, m = 5)
Analysis of Variance Table

Response: Crime
          Df  Sum Sq Mean Sq F value  Pr(>F)
M          1   55084   55084    1.26  0.2702
So         1   15370   15370    0.35  0.5575
Ed         1  905668  905668   20.72 7.7e-05 ***
Po1        1 3076033 3076033   70.38 1.8e-09 ***
Po2        1  153024  153024    3.50  0.0708 .
LF         1   61134   61134    1.40  0.2459
M.F        1  111000  111000    2.54  0.1212
Pop        1   42649   42649    0.98  0.3309
NW         1   14197   14197    0.32  0.5728
U1         1    7065    7065    0.16  0.6904
U2         1  269663  269663    6.17  0.0186 *
Wealth     1   34748   34748    0.79  0.3795
Ineq       1  547423  547423   12.52  0.0013 **
Prob       1  222620  222620    5.09  0.0312 *
Time       1   10304   10304    0.24  0.6307
Residuals 31 1354946   43708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



fold 1
Observations in test set: 9
                1    4    8   9  18      20   23    32   47
Predicted   755 1791 1362 689 844 1227.84  958 807.8  992
cvpred      658 1690 1300 617 792 1220.22  814 804.9 1077
Crime       791 1969 1555 856 929 1225.00 1216 754.0  849
CV residual 133  279  255 239 137    4.78  402 -50.9 -228

Sum of squares = 453204    Mean square = 50356    n = 9

fold 2
Observations in test set: 10
               5   13   15  17   25    34    39     40    42   46
Predicted   1167  733  903 393  606 971.5 839.3 1131.5 326.3  827
cvpred      1132  926  977 152  740 902.7 918.1 1248.5  62.3 1004
```

```
Crime       1234  511  798 539  523 923.0 826.0 1151.0 542.0  508
CV residual  102 -415 -179 387 -217  20.3 -92.1  -97.5 479.7 -496

Sum of squares = 906384    Mean square = 90638    n = 10

fold 3
Observations in test set: 10
                 2   3   11   14   16   22   28    31   33      38
Predicted   1473.7 322 1161  780 1006  657 1258 388.0  841 562.693
cvpred      1566.9 313  953  782 1129  876 1368 321.7  700 566.231
Crime       1635.0 578 1674  664  946  439 1216 373.0 1072 566.000
CV residual   68.1 265  721 -118 -183 -437 -152  51.3  372  -0.231

Sum of squares = 997216    Mean square = 99722    n = 10

fold 4
Observations in test set: 9
              19    21   26   27   29    30     36   44   45
Predicted   1146 774.9 1977  279 1287 702.7 1137.6 1121  617
cvpred      1529 802.3 1673  467 1673 629.6 1191.9 1298  702
Crime        750 742.0 1993  342 1043 696.0 1272.0 1030  455
CV residual -779 -60.3  320 -125 -630  66.4   80.1 -268 -247

Sum of squares = 1269688    Mean square = 141076    n = 9

fold 5
Observations in test set: 9
               6     7    10    12  24    35   37  41   43
Predicted    793 934.2 736.5 722.0 869 737.8  971 824 1134
cvpred       819 950.9 758.1 772.5 802 690.5 1227 891 1267
Crime        682 963.0 705.0 849.0 968 653.0  831 880  823
CV residual -137  12.1 -53.1  76.5 166 -37.5 -396 -11 -444

Sum of squares = 410109    Mean square = 45568    n = 9

Overall (Sum over all 9 folds)
   ms
85885
Warning message:
In cv.lm(mydata, model1, m = 5) :

 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

> c2 = cv.lm(mydata, model2, m = 5)
Analysis of Variance Table

Response: Crime
          Df  Sum Sq Mean Sq F value  Pr(>F)
M          1   55084   55084    1.37 0.24914
Ed         1  725967  725967   18.02 0.00013 ***
Po1        1 3173852 3173852   78.80 5.3e-11 ***
U2         1  217386  217386    5.40 0.02534 *
Ineq       1  848273  848273   21.06 4.3e-05 ***
Prob       1  249308  249308    6.19 0.01711 *
Residuals 40 1611057   40276
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



fold 1
Observations in test set: 9
                1    4    8   9  18     20   23    32   47
```

```
Predicted   810.8 1897 1354 719 800 1203.0  938 773.7  976
cvpred      762.1 1858 1282 657 672 1210.8  871 777.6  998
Crime       791.0 1969 1555 856 929 1225.0 1216 754.0  849
CV residual  28.9  111  273 199 257   14.2  345 -23.6 -149

Sum of squares = 335463    Mean square = 37274    n = 9

fold 2
Observations in test set: 10
               5   13     15    17   25   34    39   40  42   46
Predicted   1270  739 828.34 527.4  579  998 786.7 1141 369  748
cvpred      1337  842 804.73 469.3  671 1032 810.3 1187 302  839
Crime       1234  511 798.00 539.0  523  923 826.0 1151 542  508
CV residual -103 -331  -6.73  69.7 -148 -109  15.7  -36 240 -331

Sum of squares = 327423    Mean square = 32742    n = 10

fold 3
Observations in test set: 10
               2   3   11    14     16   22     28    31   33    38
Predicted   1388 386 1118 713.6 1004.4  728 1259.0 440.4  874 544.4
cvpred      1368 390 1019 711.8  985.8  767 1252.6 423.8  850 511.2
Crime       1635 578 1674 664.0  946.0  439 1216.0 373.0 1072 566.0
CV residual  267 188  655 -47.8  -39.8 -328  -36.6 -50.8  222  54.8

Sum of squares = 702726    Mean square = 70273    n = 10

fold 4
Observations in test set: 9
              19    21     26     27   29    30   36   44   45
Predicted   1221 783.3 1789.1 312.20 1495 668.0 1102 1178  622
cvpred      1316 836.4 1895.7 334.15 1693 631.2 1163 1191  612
Crime        750 742.0 1993.0 342.00 1043 696.0 1272 1030  455
CV residual -566 -94.4   97.3   7.85 -650  64.8  109 -161 -157

Sum of squares = 827924    Mean square = 91992    n = 9

fold 5
Observations in test set: 9
              6   7    10  12    24   35   37    41   43
Predicted   730 733 787.3 673 919.4  808  992 796.4 1017
cvpred      707 694 776.8 660 879.7  777 1115 812.6 1091
Crime       682 963 705.0 849 968.0  653  831 880.0  823
CV residual -25 269 -71.8 189  88.3 -124 -284  67.4 -268

Sum of squares = 294201    Mean square = 32689    n = 9

Overall (Sum over all 9 folds)
   ms
52931
Warning message:
In cv.lm(mydata, model2, m = 5) :

 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

>
> # Now compare the models using the R^2 values. From the summaries printed earlier
> # we know Model 1's R2 was 0.803 and Model 2's R2 was .766.
>
> SStot = sum((mydata$Crime - mean(mydata$Crime)) ^ 2)
>
> SSc1 = attr(c1, "ms") * nrow(mydata)
> SSc2 = attr(c2, "ms") * nrow(mydata)
```

```
>
> R2_cvm1 = 1 - SSc1 / SStot
> R2_cvm2 = 1 - SSc2 / SStot
> R2_cvm1
[1] 0.413
> R2_cvm2
[1] 0.638
>
> # So we see that the first model was overfit to the data. While the R2 of
> # model 1 was initially higher than model 2 using all of the data, by using
> # 5 fold cross validation we see that model 2 has a better fit, though it is
> # still probably over-fit as we only have a small set of data. As expected, the R2
> # of model 2 using cross validation is lower than that of the whole data set.
```