Linear Regression Modelling

Kent Ng

May 31, 2018

Using the dataset “8.2uscrimesummer2018.txt”, let’s build a linear model and use it to predict the crime rate of a new city with the following data:

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

Let’s first load the data into a dataframe. We will also set a seed value as best practice.

```
set.seed(42)
data_8<-read.table("8.2uscrimeSummer2018.txt",header=TRUE)
```

Let’s do a quick exploratory analysis on our data. Using the summary method and the boxplot, we can see roughly the distribution of our data.

```
head(data_8)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
##       Prob    Time Crime
## 1 0.084602 26.2011   791
## 2 0.029599 25.2999  1635
## 3 0.083401 24.3006   578
## 4 0.015801 29.9012  1969
## 5 0.041399 21.2998  1234
## 6 0.034201 20.9995   682
```

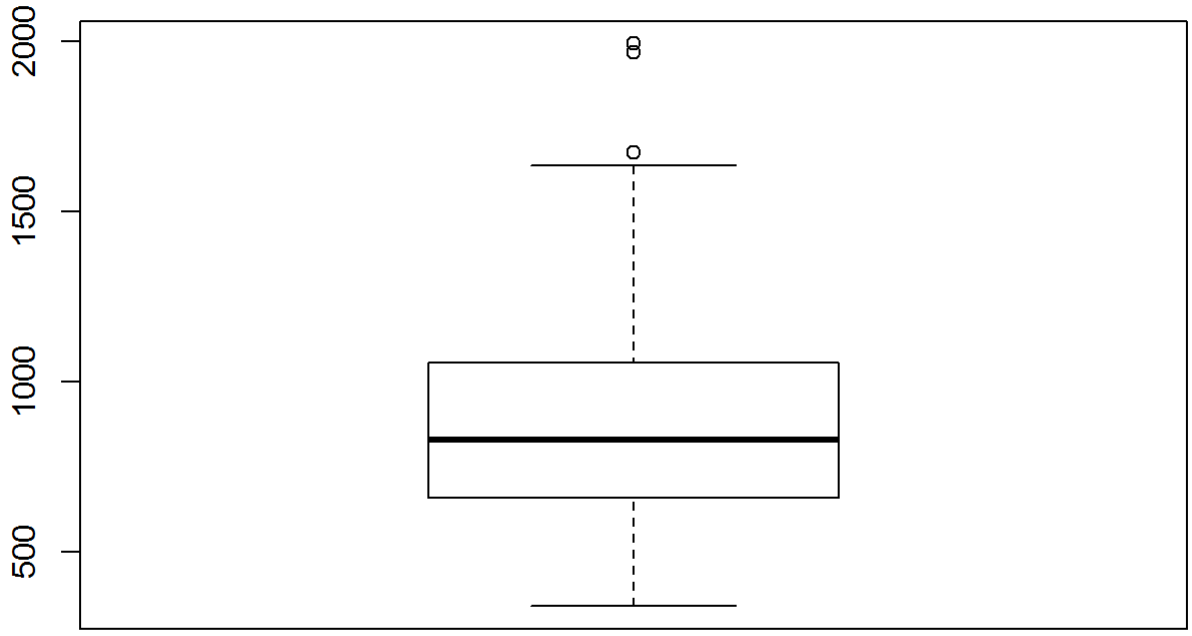
```
str(data_8)
```

```
## 'data.frame':    47 obs. of  16 variables:
##  $ M     : num  15.1 14.3 14.2 13.6 14.1 12.1 12.7 13.1 15.7 14 ...
##  $ So    : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed    : num  9.1 11.3 8.9 12.1 12.1 11 11.1 10.9 9 11.8 ...
##  $ Po1   : num  5.8 10.3 4.5 14.9 10.9 11.8 8.2 11.5 6.5 7.1 ...
##  $ Po2   : num  5.6 9.5 4.4 14.1 10.1 11.5 7.9 10.9 6.2 6.8 ...
##  $ LF    : num  0.51 0.583 0.533 0.577 0.591 0.547 0.519 0.542 0.553 0.632 ...
##  $ M.F   : num  95 101.2 96.9 99.4 98.5 ...
##  $ Pop   : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW    : num  30.1 10.2 21.9 8 3 4.4 13.9 17.9 28.6 1.5 ...
##  $ U1    : num  0.108 0.096 0.094 0.102 0.091 0.084 0.097 0.079 0.081 0.1 ...
##  $ U2    : num  4.1 3.6 3.3 3.9 2 2.9 3.8 3.5 2.8 2.4 ...
##  $ Wealth: int  3940 5570 3180 6730 5780 6890 6200 4720 4210 5260 ...
##  $ Ineq  : num  26.1 19.4 25 16.7 17.4 12.6 16.8 20.6 23.9 17.4 ...
##  $ Prob  : num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time  : num  26.2 25.3 24.3 29.9 21.3 ...
##  $ Crime : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
```

```
summary(data_8)
```

```
##        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
```

```
boxplot(data_8$Crime)
```



Next, let's build our linear regression model based on the provided data.

```
model_all<-lm(Crime~.,data_8)
```

Using the model we just developed, lets estimate the crime rate based on the given parameters in the question

```
datapoint<-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)
predict(model_all,datapoint)
```

```
##        1
## 155.4349
```

If we compare the crime rate prediction of 155.43 with our given data set, this prediction seems to be an outlier (review box plot above). As the question mentioned, because we have 15 predictors but only 47 data points, our model is likely overfitted and has a high degree of error. Although we were given 15 predictors, it is likely that not all of them are useful for our model. Let's now take a look at the P value of each coefficient. This is fourth column of the summary below. We can see that the predictors M, Ed, Ineq and Prob are signficant based on an alpha value of 0.05. We can also see that predictors Po1 and U2 barely missed the 0.05 threshold. Given that 0.05 is an arbitrary cutoff, we will also play around with predictors Po1 and U2 to determine whether they actually help our model.

One additional note: the outputted R squared value of 0.7078 is based on our training data. Thus we should not use the value to comment on the model's overall quality of fit, given that it is likely going to be too optimistic; applying the model on another (validation) set would likely yield a much lower R squared value (you will see this as we calculate the cross validated R squared value below).

```
summary(model_all)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = data_8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

Let's include only the significant predictors. As mentioned previously, since the threshold of 0.05 is arbitrary (and also because we have limited number of data points), we will also include Po1 and U2 in this simplified model

```
model_sig<-lm(Crime~M+Ed+Po1+U2+Ineq+Prob,data_8)
summary(model_sig)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = data_8)
##
## 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
```

Now let's use our new model to, once again, predict the crime rate based on the given parameters in the question.

```
predict(model_sig,datapoint)
```

```
##        1
## 1304.245
```

This time, the crime rate prediction of 1304.245 seems to be much more in line with the rest of the data. Let's now determine our model's quality of fit. We can use the cross validation for linear regression method in the DAAG (Data Analysis and Graphics Data and Functions) package

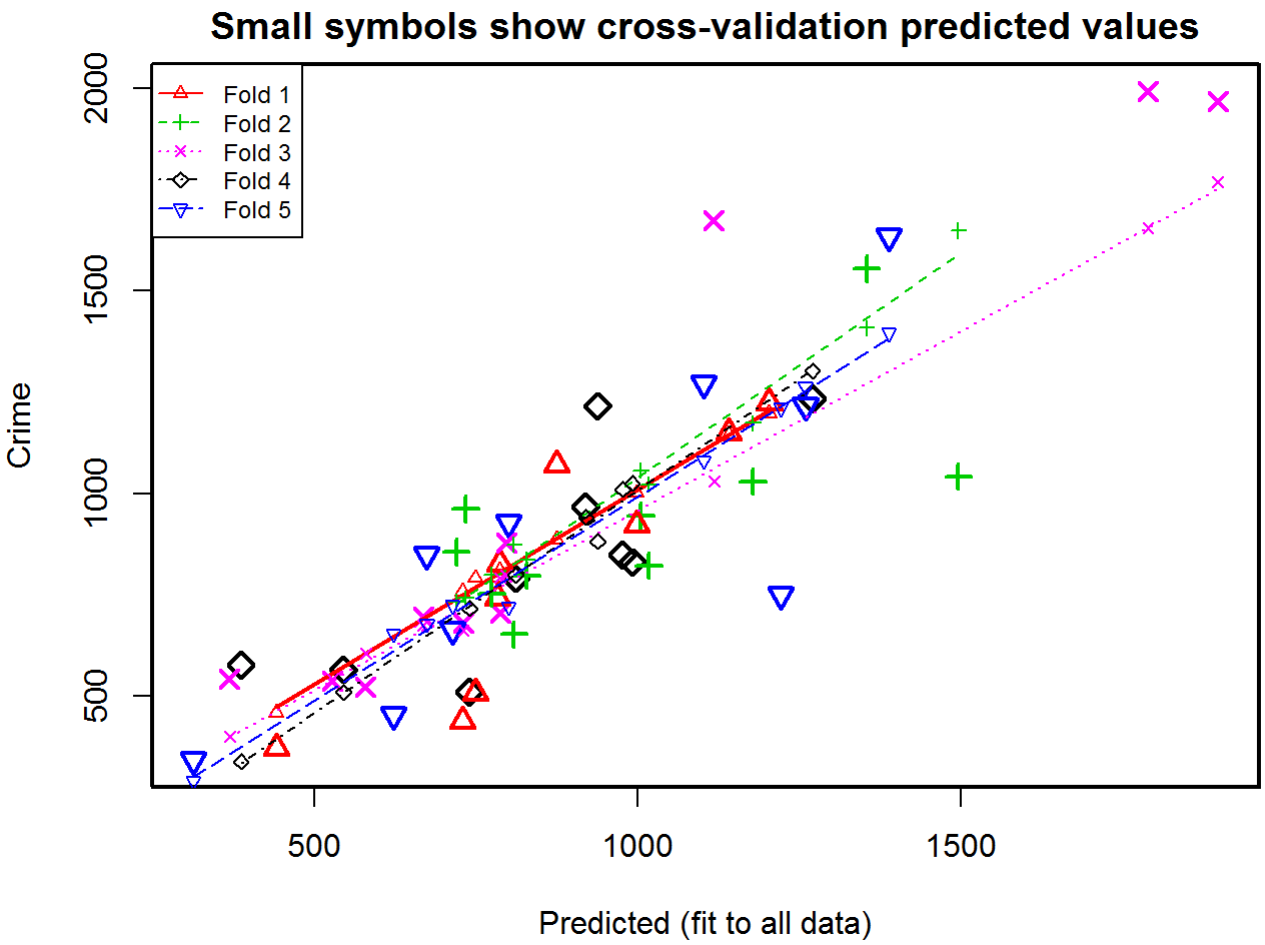
```
library(DAAG)
```

```
## Loading required package: lattice
```

```
model_sig_cv<-cv.lm(data_8, model_sig,seed=42, 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
```

```
## Warning in cv.lm(data_8, model_sig, seed = 42, 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
```



```
##
## fold 1
## Observations in test set: 9
##                 20    21   22    31   33     34    39      40   46
## Predicted   1203.0 783.3  728 440.4  874  997.5 786.7 1140.79  748
## cvpred      1198.8 793.6  759 459.5  887 1001.7 810.2 1146.01  792
## Crime       1225.0 742.0  439 373.0 1072  923.0 826.0 1151.00  508
## CV residual   26.2 -51.6 -320 -86.5  185  -78.7  15.8    4.99 -284
##
## Sum of squares = 234509    Mean square = 26057    n = 9
##
## fold 2
## Observations in test set: 10
##                7    8   9    15   16   29  32   35   43   44
## Predicted   733 1354 719 828.3 1004 1495 774  808 1017 1178
## cvpred      742 1409 733 836.9 1057 1649 800  874 1023 1175
## Crime       963 1555 856 798.0  946 1043 754  653  823 1030
## CV residual 221  146 123 -38.9 -111 -606 -46 -221 -200 -145
##
## Sum of squares = 578275    Mean square = 57827    n = 10
##
## fold 3
## Observations in test set: 10
##                 4     6    10   11     17  25   26  30    41  42
## Predicted   1897 730.3 787.3 1118 527.37 579 1789 668 796.4 369
## cvpred      1770 663.4 792.1 1031 541.22 605 1655 676 797.7 401
## Crime       1969 682.0 705.0 1674 539.00 523 1993 696 880.0 542
## CV residual  199  18.6 -87.1  643  -2.22 -82  338  20  82.3 141
##
## Sum of squares = 609041    Mean square = 60904    n = 10
##
## fold 4
## Observations in test set: 9
##                   1   3      5   13   23    24   37    38   47
## Predicted   810.83 386 1269.8  739  938 919.4  992 544.4  976
## cvpred      799.34 339 1302.7  717  882 941.3 1025 510.1 1010
## Crime       791.00 578 1234.0  511 1216 968.0  831 566.0  849
## CV residual  -8.34 239  -68.7 -206  334  26.7 -194  55.9 -161
##
## Sum of squares = 283612    Mean square = 31512    n = 9
##
## fold 5
## Observations in test set: 9
##                 2  12    14  18   19    27     28   36   45
## Predicted   1388 673 713.6 800 1221 312.2 1259.0 1102  622
## cvpred      1396 679 724.2 721 1212 291.4 1262.5 1082  655
## Crime       1635 849 664.0 929  750 342.0 1216.0 1272  455
## CV residual  239 170 -60.2 208 -462  50.6  -46.5  190 -200
##
## Sum of squares = 426429    Mean square = 47381    n = 9
##
## Overall (Sum over all 9 folds)
```

```
##    ms
## 45359
```

Therefore, the average mean squared error of the model is 45359. We can calculate the sum of squared errors, total sum of squared differences, and finally, the R squared value.

```
# Calculate Mean Squared Error
MSE_sig = attr(model_sig_cv,"ms")
MSE_sig
```

```
## [1] 45359
```

```
# Calculate Sum of Squared Errors
SSR_sig<-MSE_sig*nrow(data_8)
SSR_sig
```

```
## [1] 2131865
```

```
# Calculate Sum of Squared Differences
SSTOT_sig<-sum((data_8$Crime - mean(data_8$Crime))^2)
SSTOT_sig
```

```
## [1] 6880928
```

```
# Calculate R Squared
R2_sig<-1-(SSR_sig/SSTOT_sig)
R2_sig
```

```
## [1] 0.69
```

Therefore, the calculated overall R squared value is 0.69, meaning that 69% of the variation in the data is caused by variations in the predictors. As mentioned in the lectures, a R squared value of 0.69 (especially after cross validation) is considered quite high in many real life situations. Thus, the model_sig is likely to be an adequate (good enough) model.

For the purpose of having a comparison, however, let's build one more linear regression model using predictors with a p value of strictly 0.05 or lower (excludes Po1 and U2 in model_sig)

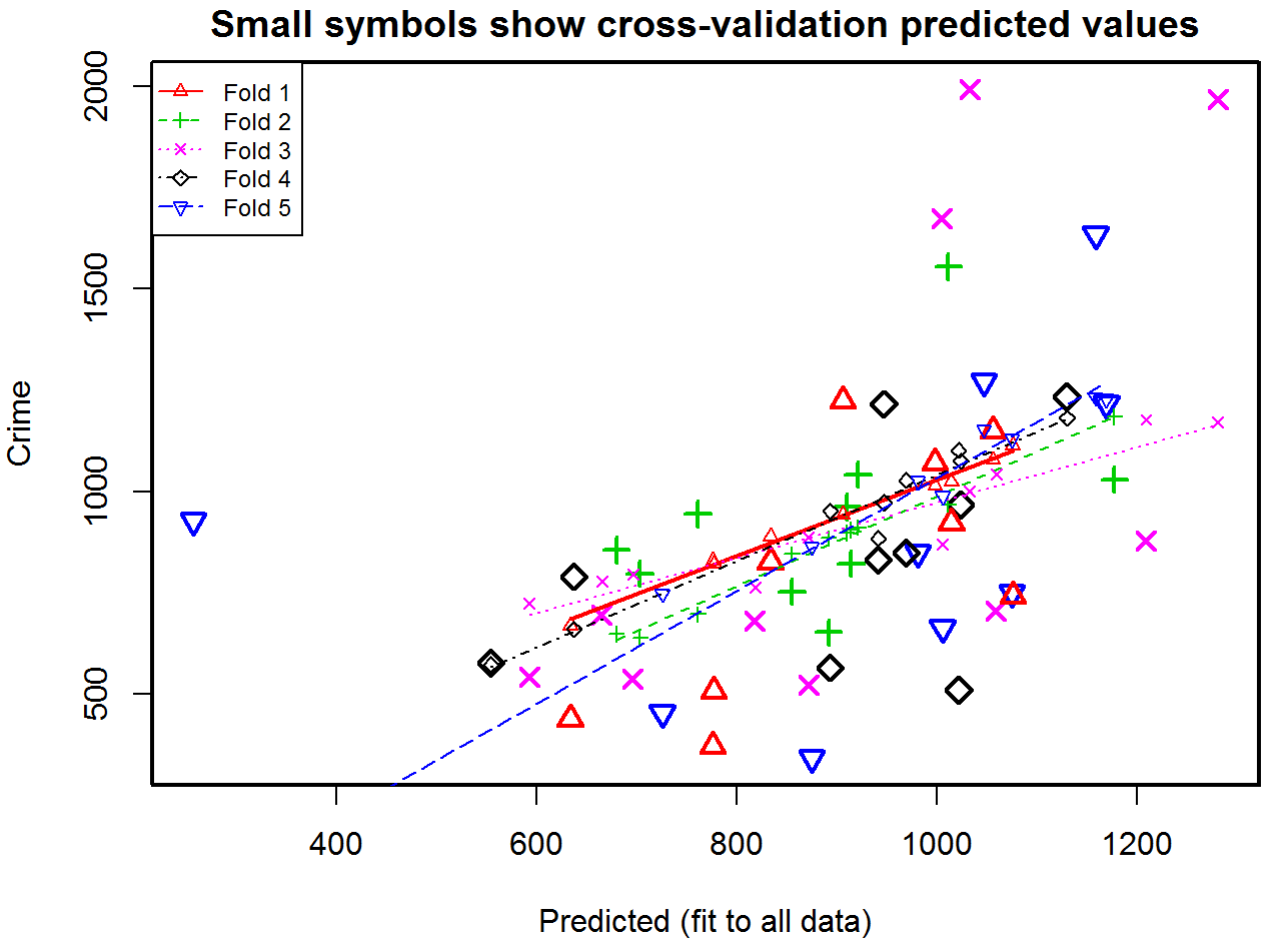
```
model_0.05<-lm(Crime~M+Ed+Ineq+Prob,data_8)
summary(model_0.05)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob, data = data_8)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -533.0 -254.0  -55.7  137.8  960.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1339.3     1247.0   -1.07   0.2889
## M               36.0       53.4    0.67   0.5042
## Ed             148.6       71.9    2.07   0.0450 *
## Ineq            26.9       22.8    1.18   0.2446
## Prob         -7331.9     2560.3   -2.86   0.0065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 348 on 42 degrees of freedom
## Multiple R-squared:  0.263,  Adjusted R-squared:  0.193
## F-statistic: 3.75 on 4 and 42 DF,  p-value: 0.0108
```

```
model_0.05_cv<-cv.lm(data_8, model_0.05,seed=42, m=5)
```

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value Pr(>F)
## M          1   55084   55084    0.46 0.5031
## Ed         1  725967  725967    6.01 0.0185 *
## Ineq       1   37674   37674    0.31 0.5794
## Prob       1  990334  990334    8.20 0.0065 **
## Residuals 42 5071868  120759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in cv.lm(data_8, model_0.05, seed = 42, 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
```


Small symbols show cross-validation predicted values
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Crime
2000
1500
1000
500
400
600
800
1000
1200
Predicted (fit to all data)

```
##
## fold 1
## Observations in test set: 9
##               20   21   22   31     33   34    39     40   46
## Predicted    906 1076  634  776  998.2 1014 833.6 1056.0  777
## cvpred       944 1115  670  831 1015.9 1024 889.9 1078.1  822
## Crime       1225  742  439  373 1072.0  923 826.0 1151.0  508
## CV residual  281 -373 -231 -458   56.1 -101 -63.9   72.9 -314
##
## Sum of squares = 602393    Mean square = 66933    n = 9
##
## fold 2
## Observations in test set: 10
##                 7    8   9  15  16   29    32   35    43   44
## Predicted   909.8 1011 680 703 761  921 855.1  892 914.1 1177
## cvpred      898.4  968 649 640 698  911 845.8  886 905.9 1185
## Crime       963.0 1555 856 798 946 1043 754.0  653 823.0 1030
## CV residual  64.6  587 207 158 248  132 -91.8 -233 -82.9 -155
##
## Sum of squares = 588546    Mean square = 58855    n = 10
##
## fold 3
## Observations in test set: 10
##                4     6   10   11   17   25   26    30   41   42
## Predicted   1281 818.4 1059 1005  696  872 1033 665.2 1209  593
## cvpred      1171 763.7 1044  870  797  888 1001 777.8 1178  724
## Crime       1969 682.0  705 1674  539  523 1993 696.0  880  542
## CV residual  798 -81.7 -339  804 -258 -365  992 -81.8 -298 -182
##
## Sum of squares = 2716554    Mean square = 271655    n = 10
##
## fold 4
## Observations in test set: 9
##               1      3      5   13   23   24    37   38   47
## Predicted   637 554.43 1130.1 1022  947 1024 941.5  893  969
## cvpred      661 573.21 1181.8 1102  973 1076 884.4  953 1028
## Crime       791 578.00 1234.0  511 1216  968 831.0  566  849
## CV residual 130   4.79   52.2 -591  243 -108 -53.4 -387 -179
##
## Sum of squares = 624245    Mean square = 69361    n = 9
##
## fold 5
## Observations in test set: 9
##                2   12   14  18   19   27     28   36   45
## Predicted   1159  981 1006 257 1075  875 1169.5 1047  726
## cvpred      1232 1025  988 -50 1130  864 1228.6 1154  748
## Crime       1635  849  664 929  750  342 1216.0 1272  455
## CV residual  403 -176 -324 979 -380 -522  -12.6  118 -293
##
## Sum of squares = 1773888    Mean square = 197099    n = 9
##
## Overall (Sum over all 9 folds)
```

```
##     ms
## 134162
```

We can already see that the mean squared error value is greater than that of our previous model (with predictors Po1 and U2 included). This means that the R squared value will be much lower (shown below), and that model_sig is better than model_0.05

model_sig is a better model than model_0.05.

```
# Calculate Mean Squared Error
MSE_0.05 = attr(model_0.05_cv,"ms")
MSE_0.05
```

```
## [1] 134162
```

```
# Calculate Sum of Squared Errors
SSR_0.05<-MSE_0.05*nrow(data_8)
SSR_0.05
```

```
## [1] 6305625
```

```
# Sum of Sqaured Differences is the same as the previous model
SSTOT_sig
```

```
## [1] 6880928
```

```
# Calculate R Squared
R2_0.05<-1-(SSR_0.05/SSTOT_sig)
R2_0.05
```

```
## [1] 0.0836
```