HW 5

Jeff Tilton

9/22/2018

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Response

Linear regression can be used to determine how systems work and to predict future outcomes. There are many situations when this type of model would be appropriate. A company could use it to determine if there is work place bias in how they promote staff. Often times companies keep data on their employees these data could be used to create a linear regression model and the coefficients would determine if race or sex plays a part in a person's promotoion. Some possible predictors are:

1. Race
2. Sex
3. Hours worked (including overtime)
4. Educational achievement
5. Training events attended

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

Response

Goals

1. Use regression to predict the observed crime rate in a city
2. Display model, software output and quality of fit

Method

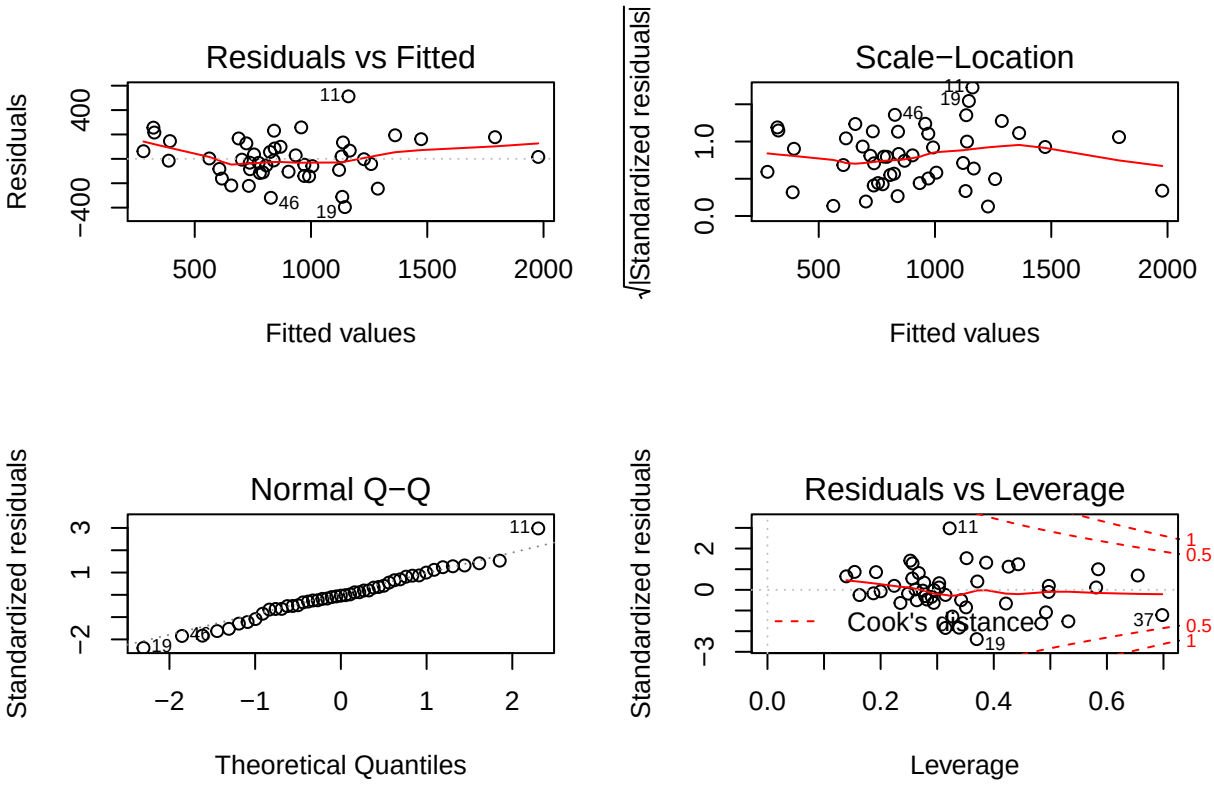
I will first use all of the predictors in the data to fit a regression model using the lm function. Next, I will remove any predictors that have a large p value and create a new model.

Output for model using all predictors

```
##
## Call:
## lm(formula = Crime ~ ., data = data)
```

```
##
## 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 *
## So1         -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
```

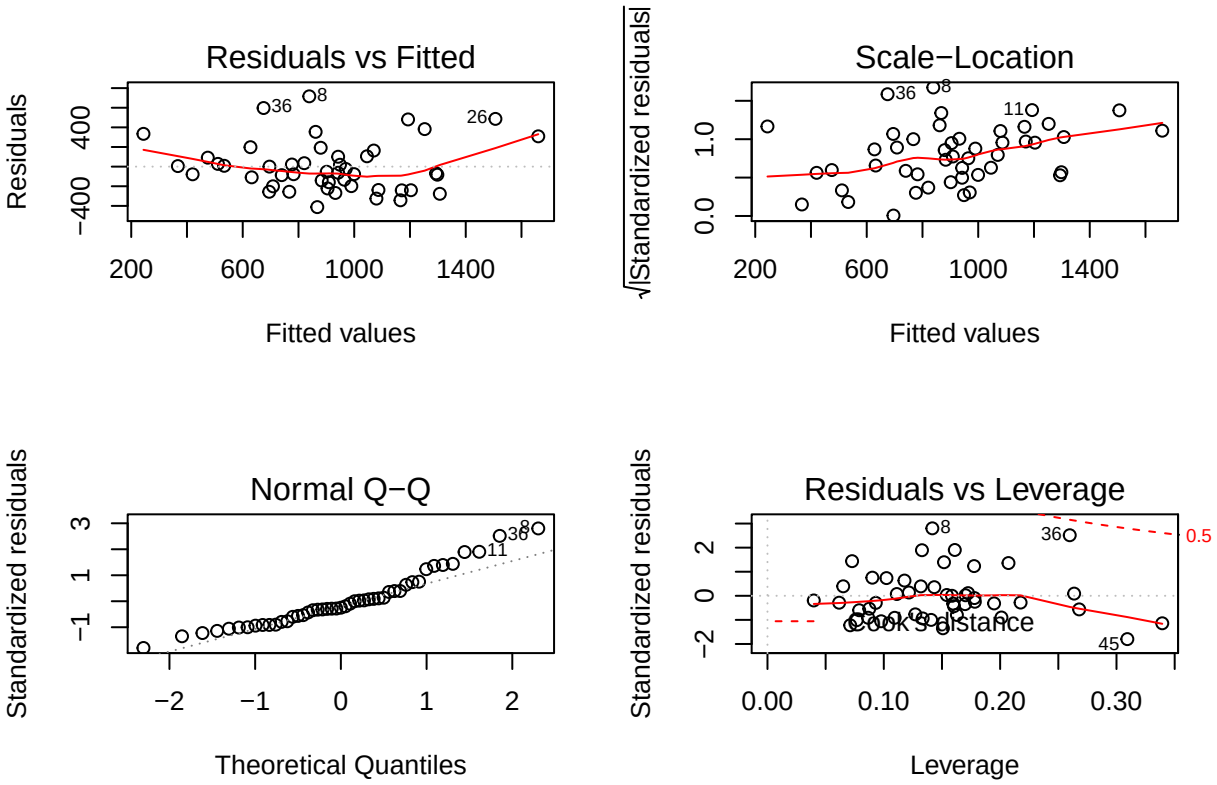
Diagnostic plots for all predictors



New model output

```
##
## Call:
## lm(formula = Crime ~ M + Ed + U2 + Ineq + Wealth + NW, data = raw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -412.45 -199.57  -62.05  103.01  716.21
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.935e+03  1.427e+03  -4.860 1.85e-05 ***
## M            9.366e+01  4.900e+01   1.911  0.06315 .
## Ed           1.826e+02  6.577e+01   2.777  0.00832 **
## U2           1.004e+02  5.652e+01   1.777  0.08312 .
## Ineq         8.567e+01  2.475e+01   3.461  0.00129 **
## Wealth       4.758e-01  9.961e-02   4.777 2.40e-05 ***
## NW           1.081e+01  5.988e+00   1.806  0.07847 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276 on 40 degrees of freedom
## Multiple R-squared:  0.5573, Adjusted R-squared:  0.4909
## F-statistic: 8.391 on 6 and 40 DF,  p-value: 6.476e-06
```

Diagnostic plots for new model



Discussion

The model using all predictors had a much higher Adjusted R-squared value, 0.7078 compared to 0.4909 as well as a p-value a whole order of magnitude smaller 3.539e-07 compared to 6.476e-06, the model diagnostic plots look similar. Although the quality of fit indicators suggest the first model may be better, the homework instructions suggest that this is a case of overfitting.

I found a cross validation function to see what the sum of mean squared error is for the two models to test both models' fit. The results from cross validation below show that the first model has an overall mean squared error if 278973 compared to the simpler models 90926. Model performance for the first model using all predictors was significantly worse although it had better quality of fit indicators.

Cross Validation

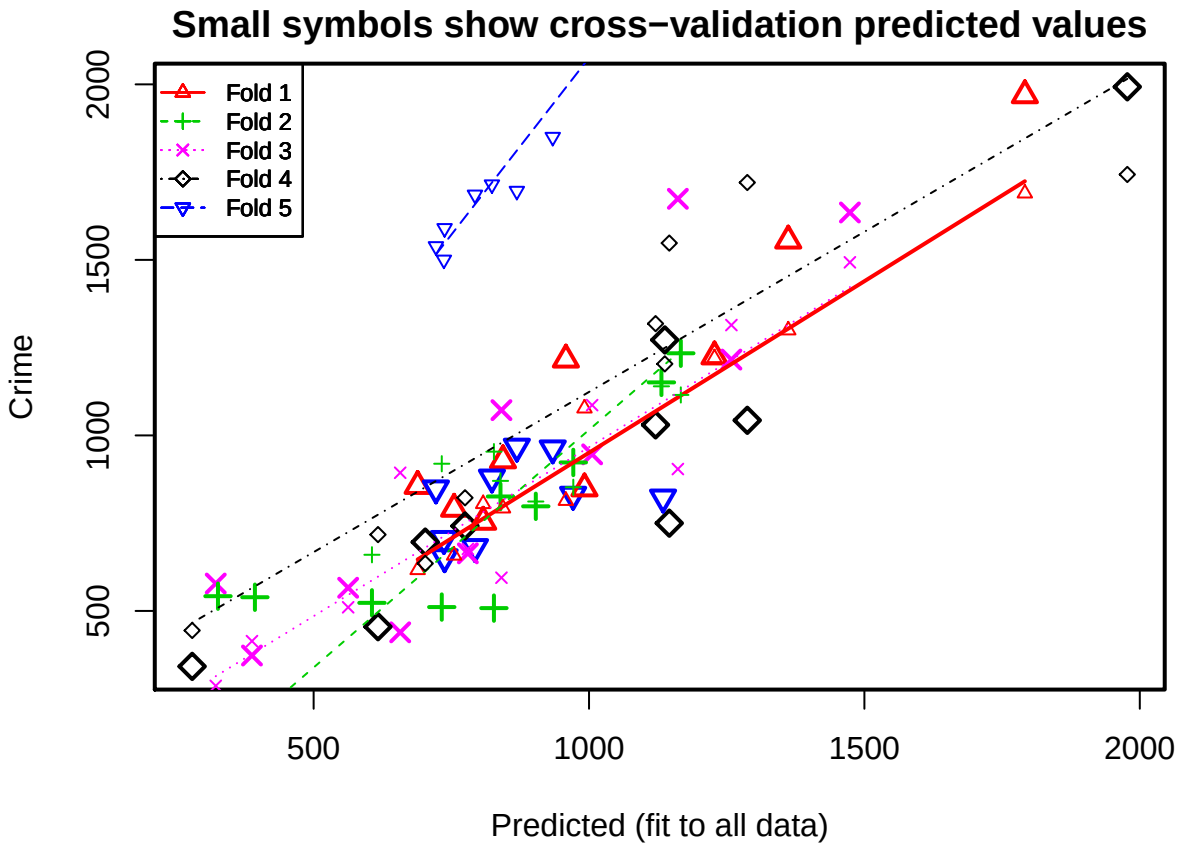
Model 1

```
## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from
## a rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from
## a rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from
## a rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from
## a rank-deficient fit may be misleading
```

```
## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from
## a rank-deficient fit may be misleading
```

```
## 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
```

```
## Warning in cv.lm(data = raw, form.lm = form.lm.1, 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
```

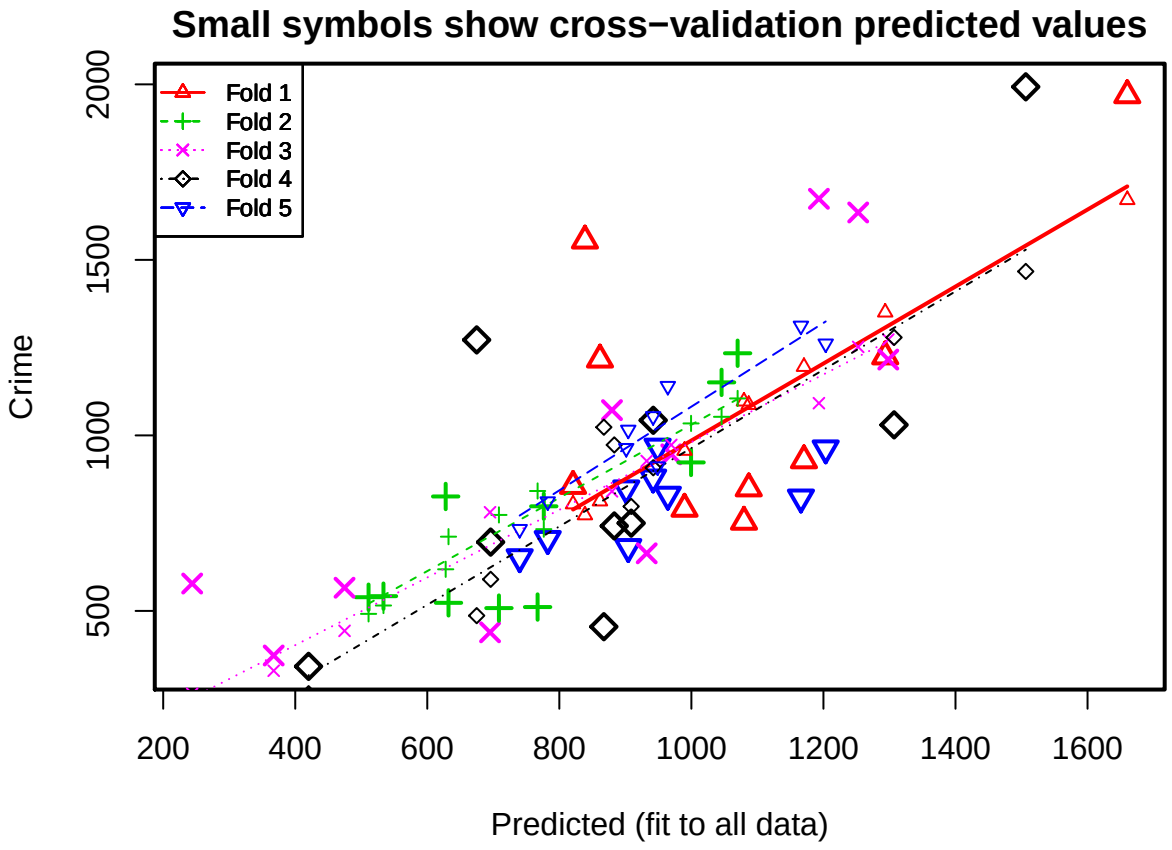



```
##
## fold 1
## Observations in test set: 9
##                1    4    8   9  18      20   23    32   47
## Predicted    755 1791 1362 689 844 1227.84  958 807.8  992
## cvpred       658 1690 1300 617 792 1220.22  814 804.9 1077
## Crime        791 1969 1555 856 929 1225.00 1216 754.0  849
## CV 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  733 903.4 393  606 971.5 839.3 1131.5 326  827
## cvpred      1115  919 811.7  68  660 852.3 870.5 1139.8 -79  954
## Crime       1234  511 798.0 539  523 923.0 826.0 1151.0 542  508
## CV residual  119 -408 -13.7 471 -137  70.7 -44.5   11.2 621 -446
##
## Sum of squares = 1013064    Mean square = 101306    n = 10
##
## fold 3
## Observations in test set: 10
##                2   3   11    14   16   22     28    31   33    38
## Predicted   1474 322 1161 780.0 1006  657 1258.5 388.0  841 562.7
## cvpred      1493 287  904 676.5 1086  894 1314.3 413.9  595 510.4
## Crime       1635 578 1674 664.0  946  439 1216.0 373.0 1072 566.0
```

```
## CV residual  142 291  770 -12.5 -140 -455  -98.3 -40.9  477  55.6
##
## Sum of squares = 1166539    Mean square = 116654    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      1548 822.3 1743  444 1720 635.2 1203.8 1318  717
## Crime        750 742.0 1993  342 1043 696.0 1272.0 1030  455
## CV residual -798 -80.3  250 -102 -677  60.8   68.2 -288 -262
##
## Sum of squares = 1335094    Mean square = 148344    n = 9
##
## fold 5
## Observations in test set: 9
##                 6    7   10   12   24   35    37   41    43
## Predicted     793  934  737  722  869  738   971  824  1134
## cvpred       1686 1850 1500 1539 1696 1590  2217 1715  2312
## Crime         682  963  705  849  968  653   831  880   823
## CV residual -1004 -887 -795 -690 -728 -937 -1386 -835 -1489
##
## Sum of squares = 9143814    Mean square = 1015979    n = 9
##
## Overall (Sum over all 9 folds)
##     ms
## 278973
```

Model 2

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    0.72  0.4001
## Ed         1  725967  725967    9.53  0.0037 **
## U2         1  736262  736262    9.67  0.0034 **
## Ineq       1   63813   63813    0.84  0.3655
## Wealth     1 2005043 2005043   26.33 7.8e-06 ***
## NW         1  248363  248363    3.26  0.0785 .
## Residuals 40 3046395   76160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Warning in cv.lm(data = raw, form.lm = form.lm.2, 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
##                 1    4    8   9   18   20   23   32   47
## Predicted     990 1660  839 821 1171 1294  862 1080 1087
## cvpred        957 1671  773 804 1196 1351  812 1097 1087
## Crime         791 1969 1555 856  929 1225 1216  754  849
## CV residual -166  298  782  52 -267 -126  404 -343 -238
##
## Sum of squares = 1154878    Mean square = 128320    n = 9
##
## fold 2
## Observations in test set: 10
##                  5   13    15    17   25   34  39     40    42   46
## Predicted   1070  767 776.4 511.0  632 1000 628 1045.7 533.6  709
## cvpred      1105  842 732.8 491.6  711 1034 619 1052.9 515.4  773
## Crime       1234  511 798.0 539.0  523  923 826 1151.0 542.0  508
## CV residual  129 -331  65.2  47.4 -188 -111 207   98.1  26.6 -265
##
## Sum of squares = 304064    Mean square = 30406    n = 10
##
## fold 3
## Observations in test set: 10
##                  2   3   11   14    16   22     28    31   33  38
## Predicted   1253 244 1193  932 969.5  695 1298.2 367.4  880 475
## cvpred      1252 264 1092  927 972.9  781 1272.9 329.5  841 443
## Crime       1635 578 1674  664 946.0  439 1216.0 373.0 1072 566
```

```
## CV residual  383 314  582 -263 -26.9 -342  -56.9  43.5  231 123
##
## Sum of squares = 844945    Mean square = 84495    n = 10
##
## fold 4
## Observations in test set: 9
##                19   21   26    27   29  30   36   44   45
## Predicted   908.9  883 1506 420.2  942 696  675 1307  867
## cvpred      797.6  973 1467 264.2  908 590  487 1279 1023
## Crime       750.0  742 1993 342.0 1043 696 1272 1030  455
## CV residual -47.6 -231  526  77.8  135 106  785 -249 -568
##
## Sum of squares = 1369735    Mean square = 152193    n = 9
##
## fold 5
## Observations in test set: 9
##                6    7   10   12    24    35   37   41   43
## Predicted    904 1203  782  901 948.9 739.9  964  942 1166
## cvpred      1016 1261  812  964 911.2 732.9 1140 1054 1312
## Crime        682  963  705  849 968.0 653.0  831  880  823
## CV residual -334 -298 -107 -115  56.8 -79.9 -309 -174 -489
##
## Sum of squares = 6e+05    Mean square = 66653    n = 9
##
## Overall (Sum over all 9 folds)
##    ms
## 90926
```