Homework 5

Richard Albright
ISYE6501
Spring 2018

2/9/2019

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.

I work for a FinTech Company that analysis insider trading data. I have used linear regression in the past to build out a company score (ranging from 0 to 10) from various insider and company predictors. The score represents how likely the company will outperform its sector peers. The following are some of the predictors included: The sum of insider purchases and sales over a 30-day period. The length of time in days since the last purchase or sale. The percentile rank of the company's current 30-day sum of purchases and sales compared to its past quarterly average. The market cap of the company. The percentile rank of the company's 30-day purchases and sales compared to the sector's quarterly averages.

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (fileuscrime.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:

Read in the CSV

```
data <-
  read.table(
    "/Users/ralbright/Dropbox/ISYE6501/week3/homework/uscrime.txt",
    header=TRUE,
    sep="\t"
  )
```

Head:

```
table <- xtable(head(data))
print(table, type='latex', comment=FALSE, scalebox='0.75')
```

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
1	15.10	1	9.10	5.80	5.60	0.51	95.00	33	30.10	0.11	4.10	3940	26.10	0.08	26.20	791
2	14.30	0	11.30	10.30	9.50	0.58	101.20	13	10.20	0.10	3.60	5570	19.40	0.03	25.30	1635
3	14.20	1	8.90	4.50	4.40	0.53	96.90	18	21.90	0.09	3.30	3180	25.00	0.08	24.30	578
4	13.60	0	12.10	14.90	14.10	0.58	99.40	157	8.00	0.10	3.90	6730	16.70	0.02	29.90	1969
5	14.10	0	12.10	10.90	10.10	0.59	98.50	18	3.00	0.09	2.00	5780	17.40	0.04	21.30	1234
6	12.10	0	11.00	11.80	11.50	0.55	96.40	25	4.40	0.08	2.90	6890	12.60	0.03	21.00	682

Tail:

```
table <- xtable(tail(data))
print(table, type='latex', comment=FALSE, scalebox='0.75')
```

Summary:

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
42	14.10	0	10.90	5.60	5.40	0.52	96.80	4	0.20	0.11	3.70	4890	17.00	0.09	12.20	542
43	16.20	1	9.90	7.50	7.00	0.52	99.60	40	20.80	0.07	2.70	4960	22.40	0.05	32.00	823
44	13.60	0	12.10	9.50	9.60	0.57	101.20	29	3.60	0.11	3.70	6220	16.20	0.03	30.00	1030
45	13.90	1	8.80	4.60	4.10	0.48	96.80	19	4.90	0.14	5.30	4570	24.90	0.06	32.60	455
46	12.60	0	10.40	10.60	9.70	0.60	98.90	40	2.40	0.08	2.50	5930	17.10	0.05	16.70	508
47	13.00	0	12.10	9.00	9.10	0.62	104.90	3	2.20	0.11	4.00	5880	16.00	0.05	16.10	849

```r
table <- xtable(summary(data))
print(table, type='latex', comment=FALSE, scalebox='0.4')
```

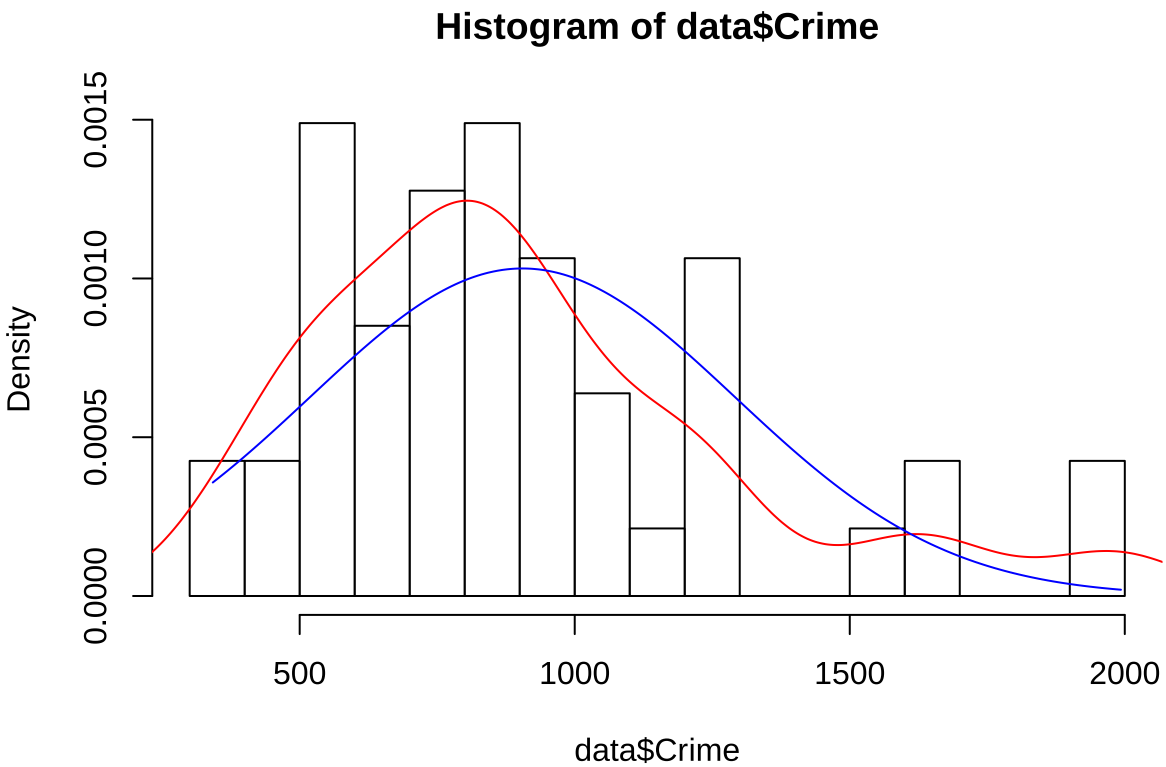
	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
X	Min. :11.90	Min. :0.0000	Min. : 8.70	Min. : 4.50	Min. : 4.100	Min. :0.4800	Min. : 93.40	Min. : 3.00	Min. : 0.20	Min. :0.07000	Min. :2.000	Min. :2880	Min. :12.60	Min. :0.00690	Min. :12.20	Min. : 342.0
X.1	1st Qu.:13.00	1st Qu.:0.0000	1st Qu.: 9.75	1st Qu.: 6.25	1st Qu.: 5.850	1st Qu.:0.5305	1st Qu.: 96.45	1st Qu.: 10.00	1st Qu.: 2.40	1st Qu.:0.08050	1st Qu.:2.750	1st Qu.:4595	1st Qu.:16.55	1st Qu.:0.03270	1st Qu.:21.60	1st Qu.: 658.5
X.2	Median :13.60	Median :0.0000	Median :10.80	Median : 7.80	Median : 7.300	Median :0.5600	Median : 97.70	Median : 25.00	Median : 7.60	Median :0.09200	Median :3.400	Median :5370	Median :17.60	Median :0.04210	Median :25.80	Median : 831.0
X.3	Mean :13.86	Mean :0.3404	Mean :10.56	Mean : 8.50	Mean : 8.023	Mean :0.5612	Mean : 98.30	Mean : 36.62	Mean :10.11	Mean :0.09547	Mean :3.398	Mean :5254	Mean :19.40	Mean :0.04709	Mean :26.60	Mean : 905.1
X.4	3rd Qu.:14.60	3rd Qu.:1.0000	3rd Qu.:11.45	3rd Qu.:10.45	3rd Qu.: 9.700	3rd Qu.:0.5930	3rd Qu.: 99.20	3rd Qu.: 41.50	3rd Qu.:13.25	3rd Qu.:0.10400	3rd Qu.:3.850	3rd Qu.:5915	3rd Qu.:22.75	3rd Qu.:0.05445	3rd Qu.:30.45	3rd Qu.:1057.5
X.5	Max. :17.70	Max. :1.0000	Max. :12.20	Max. :16.60	Max. :15.700	Max. :0.6410	Max. :107.10	Max. :168.00	Max. :42.30	Max. :0.14200	Max. :5.800	Max. :6890	Max. :27.60	Max. :0.11980	Max. :44.00	Max. :1993.0

Example analysis from http://www.statsci.org/data/general/uscrime.html

Testing our data set for outliers using grubbs.test

Lets 1st plot a histogram of our Crime Response variable vs its density and a overlay of the normal distribution.

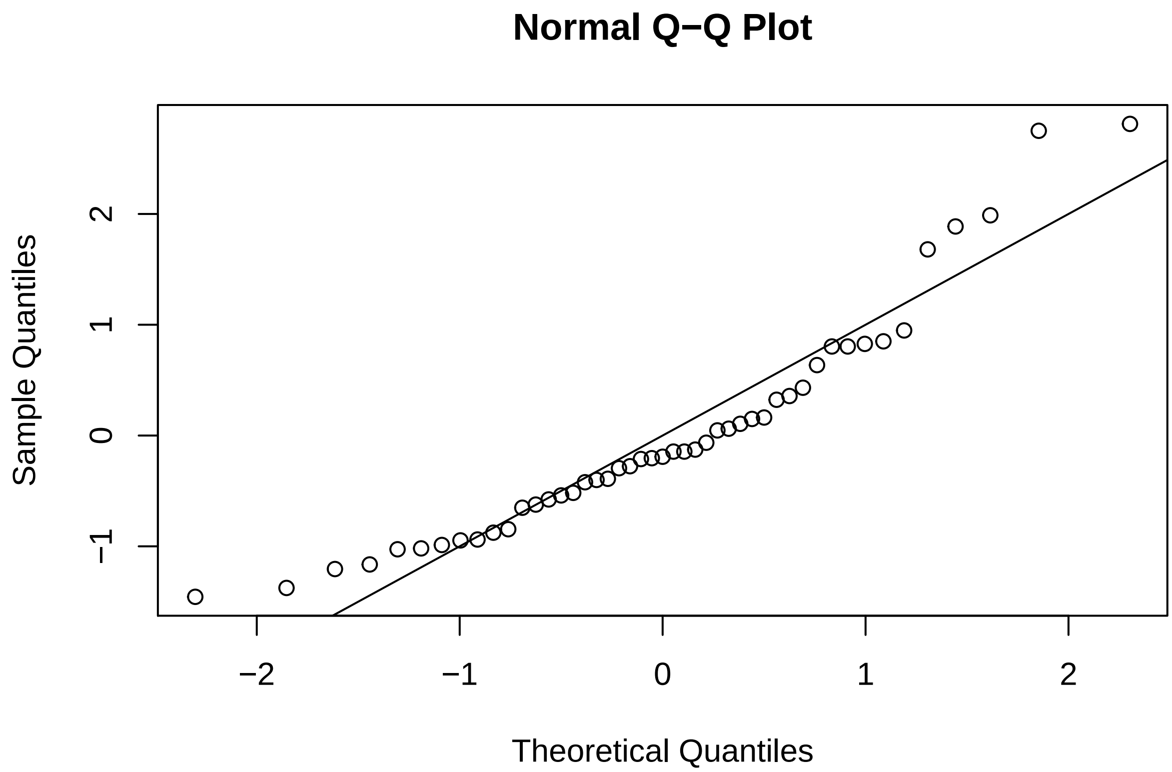
```r
hist(data$Crime, freq=F, breaks=12)
lines(density(data$Crime), col="red")
lines(seq(min(data$Crime), max(data$Crime)), dnorm(seq(min(data$Crime), max(data$Crime)),mean(data$Crim
```



The left tail seems to indicate there may be some outliers in our data set.

The plot of the scaled Crime Response Variable using qqnorm also looks like.

```
scaled_crime = scale(data$Crime)
qqnorm(scaled_crime)
abline(0,1)
```



Which seems to indicate that there may outliers in both tails.

Lets take a look at a box plot of our Crime response variable as well.

```
boxplot(data$Crime, main="Crime", boxwex=0.1)
```

```
possible_outliers <- boxplot.stats(data$Crime)$out
possible_outliers
```

```
## [1] 1969 1674 1993
```

The boxplot points to possible outliers in the upper tail. Output from boxplot.stats indicates that the 3 possible outliers are 1969, 1674, & 1993. We will now use the grubbs.test function to test for the outliers from the data set.

We will use the 1st 2 tests of the The grubbs.test function below (taken directly from the R Documentation).

First test (10) is used to detect if the sample dataset contains one outlier, statistically different than the other values. Test is based by calculating score of this outlier G (outlier minus mean and divided by sd) and comparing it to appropriate critical values. Alternative method is calculating ratio of variances of two datasets - full dataset and dataset without outlier. The obtained value called U is bound with G by simple formula.

Second test (11) is used to check if lowest and highest value are two outliers on opposite tails of sample. It is based on calculation of ratio of range to standard deviation of the sample.

We will loop through the 1st two test types on the Crime column.

```
tests <- c(10, 11)
for(test in tests) {
  for(truth in c(TRUE,FALSE)) {
    gtest <- grubbs.test(as.vector(data$Crime), type=test, opposite=truth)
    print(paste('Grubbs Test Type:', test, collapse=' '))
    print(gtest)
  }
}
```

```
## [1] "Grubbs Test Type: 10"
##
##  Grubbs test for one outlier
##
## data:  as.vector(data$Crime)
## G = 1.45590, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
##
## [1] "Grubbs Test Type: 10"
##
##  Grubbs test for one outlier
##
## data:  as.vector(data$Crime)
## G = 2.81290, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
##
## [1] "Grubbs Test Type: 11"
##
##  Grubbs test for two opposite outliers
##
## data:  as.vector(data$Crime)
## G = 4.26880, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
##
## [1] "Grubbs Test Type: 11"
##
```

```
##  Grubbs test for two opposite outliers
##
## data:  as.vector(data$Crime)
## G = 4.26880, U = 0.78103, p-value = 1
## alternative hypothesis: 342 and 1993 are outliers
```

Using a 95% confidence interval, We accept the null hypothesis that there are not any outliers in our Crime reponse variable.

We will perform a linear regression using the lm() function using the last column Crime vs its predictor columns.

```
lm.crime <- lm(Crime~., data=data, names=names(data))
summary(lm.crime,correlation=FALSE)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = data, names = names(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) -5984.28760  1628.31837  -3.675 0.000893 ***
## M              87.83017    41.71387   2.106 0.043443 *
## So             -3.80345   148.75514  -0.026 0.979765
## Ed            188.32431    62.08838   3.033 0.004861 **
## Po1           192.80434   106.10968   1.817 0.078892 .
## Po2          -109.42193   117.47754  -0.931 0.358830
## LF           -663.82615  1469.72882  -0.452 0.654654
## M.F            17.40686    20.35384   0.855 0.398995
## Pop            -0.73301     1.28956  -0.568 0.573845
## NW              4.20446     6.48089   0.649 0.521279
## U1          -5827.10272  4210.28904  -1.384 0.176238
## U2            167.79967    82.33596   2.038 0.050161 .
## Wealth          0.09617     0.10367   0.928 0.360754
## Ineq           70.67210    22.71652   3.111 0.003983 **
## Prob        -4855.26582  2272.37462  -2.137 0.040627 *
## Time           -3.47902     7.16528  -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: 0.0000003539
```

The R-squared and adjusted R-squared from our model fitting the entire data set is 0.8030868 and 0.7078062.

Lets calculate the AIC and BIC of our initial model.

```
aic1 = AIC(lm.crime)
aic1
```

```
## [1] 650.0291
```

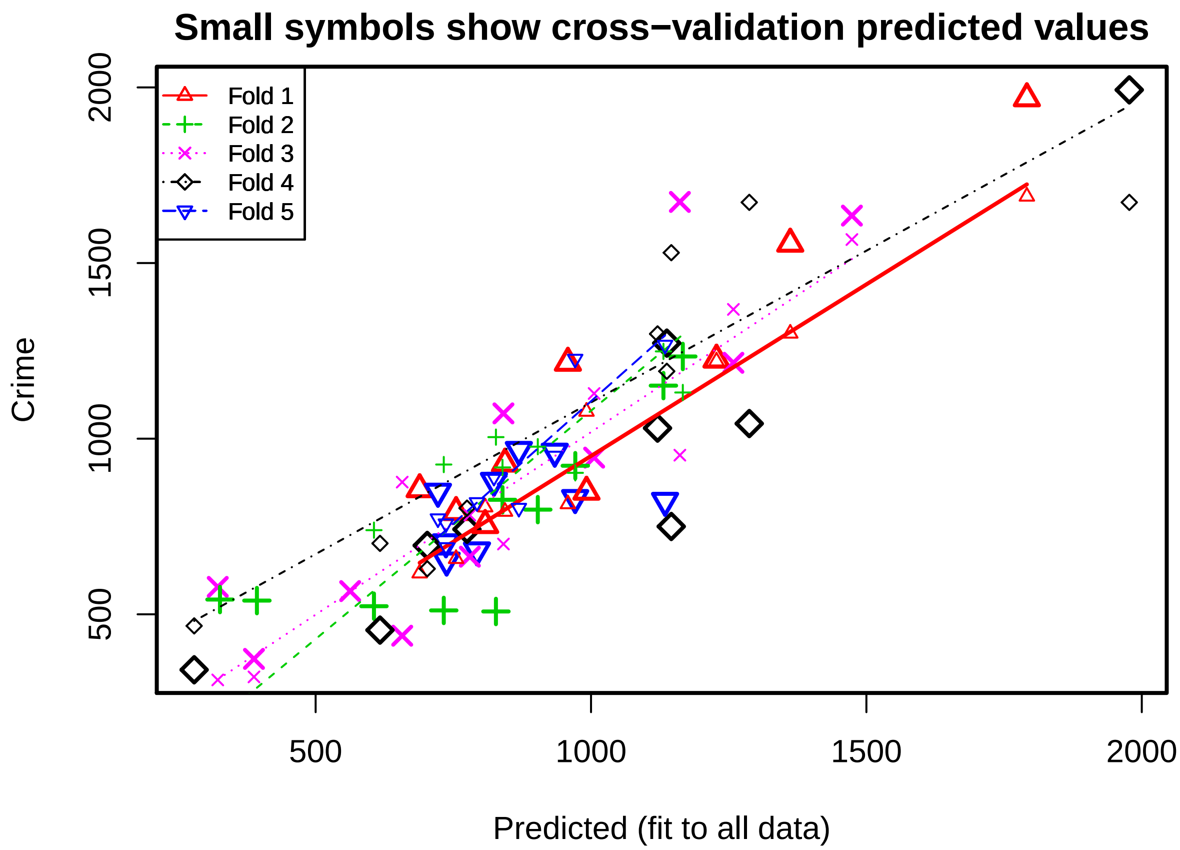
```r
bic1= BIC(lm.crime)
bic1
```

```
## [1] 681.4816
```

We'll then perform a K-Fold cross validation on our initial model using 5 folds.

```r
lm.crime.cv <- cv.lm(data, lm.crime, 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 0.0000772205 ***
## Po1        1 3076033 3076033   70.38 0.0000000018 ***
## 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
```

Then let's calculate our r^2 for our K-fold cross validated model.

```
sse1 = attr(lm.crime.cv, 'ms') * nrow(data)
sst1 = sum((data$Crime - mean(data$Crime)) ^ 2)
rsquared1 = 1 - sse1/sst1
rsquared1
```

```
## [1] 0.413
```

We find that the best predictors after performing a linear regression are M, Ed, Po1, U2, Ineq, and Prob. Our initial model's adjusted R-squared accounts for approximately 41.336% of the variance of the data set.

The leaps functions is an all subsets regression function that attempts to find the best predictors for use in a linear regression model. This can be used as an alternative to a stepwise AIC (which does stepwise regression). We can then run our predictors through the leaps functions to verify if in fact our predictors are the best ones to use (Information about leaps here: http://www2.hawaii.edu/~taylor/z632/Rbestsubsets.pdf). We want to find the combination of number of p predictors is closest in value to Mallows C_p Statistic ($p=C_p$) (https://en.wikipedia.org/wiki/Mallows's_Cp).

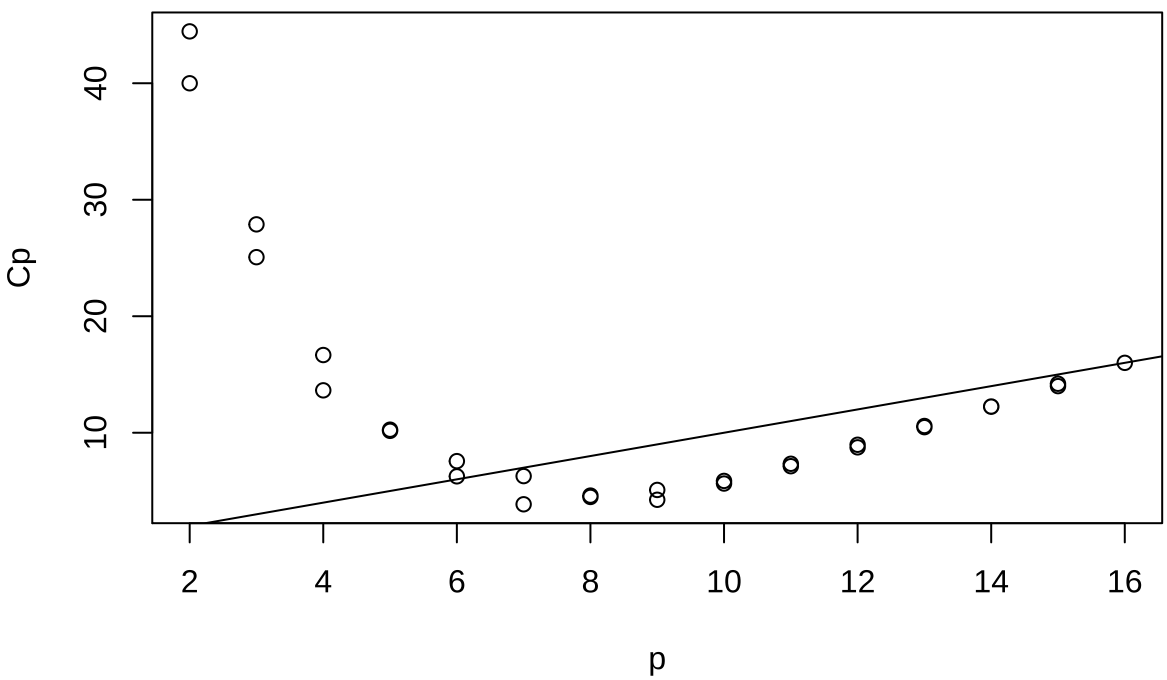
```
leaps.crime <- leaps(data[,1:15],data$Crime,nbest=2, names=names(data[,1:15]))
```

```
leaps.tab <- data.frame(p=leaps.crime$size,Cp=leaps.crime$Cp)
round(leaps.tab,2)
```

```
##     p    Cp
## 1   2 40.00
## 2   2 44.45
## 3   3 25.07
## 4   3 27.89
```

```
## 5   4 13.64
## 6   4 16.67
## 7   5 10.16
## 8   5 10.26
## 9   6  6.26
## 10  6  7.56
## 11  7  3.86
## 12  7  6.28
## 13  8  4.49
## 14  8  4.61
## 15  9  4.24
## 16  9  5.09
## 17 10  5.64
## 18 10  5.86
## 19 11  7.13
## 20 11  7.34
## 21 12  8.75
## 22 12  8.97
## 23 13 10.48
## 24 13 10.58
## 25 14 12.24
## 26 14 12.25
## 27 15 14.00
## 28 15 14.20
## 29 16 16.00
```

```
plot(leaps.tab)
abline(0,1)
```



We can see from the chart that using 6 predictors gives you the best linear regression model (The 1st point where the AB line crosses a scatter point from left to right). This agrees with what was identified as significant in our initial models K-fold cross validation. Now let's generate a linear regression model using only these factors as identified as significant in our initial K-fold cross validated lm model.

```r
lm.crime2 <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob,data=data)
summary(lm.crime2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = data)
##
## 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 0.00000171527 ***
## M              105.0       33.3    3.15        0.0031 **
## Ed             196.5       44.8    4.39 0.00008072016 ***
## Po1            115.0       13.8    8.36 0.00000000026 ***
## U2              89.4       40.9    2.18        0.0348 *
## Ineq            67.7       13.9    4.85 0.00001879377 ***
## 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: 0.0000000000342
```

Let's calculate the AIC and BIC of our improved model.

```r
aic2 = AIC(lm.crime2)
aic2
```

```
## [1] 640
```

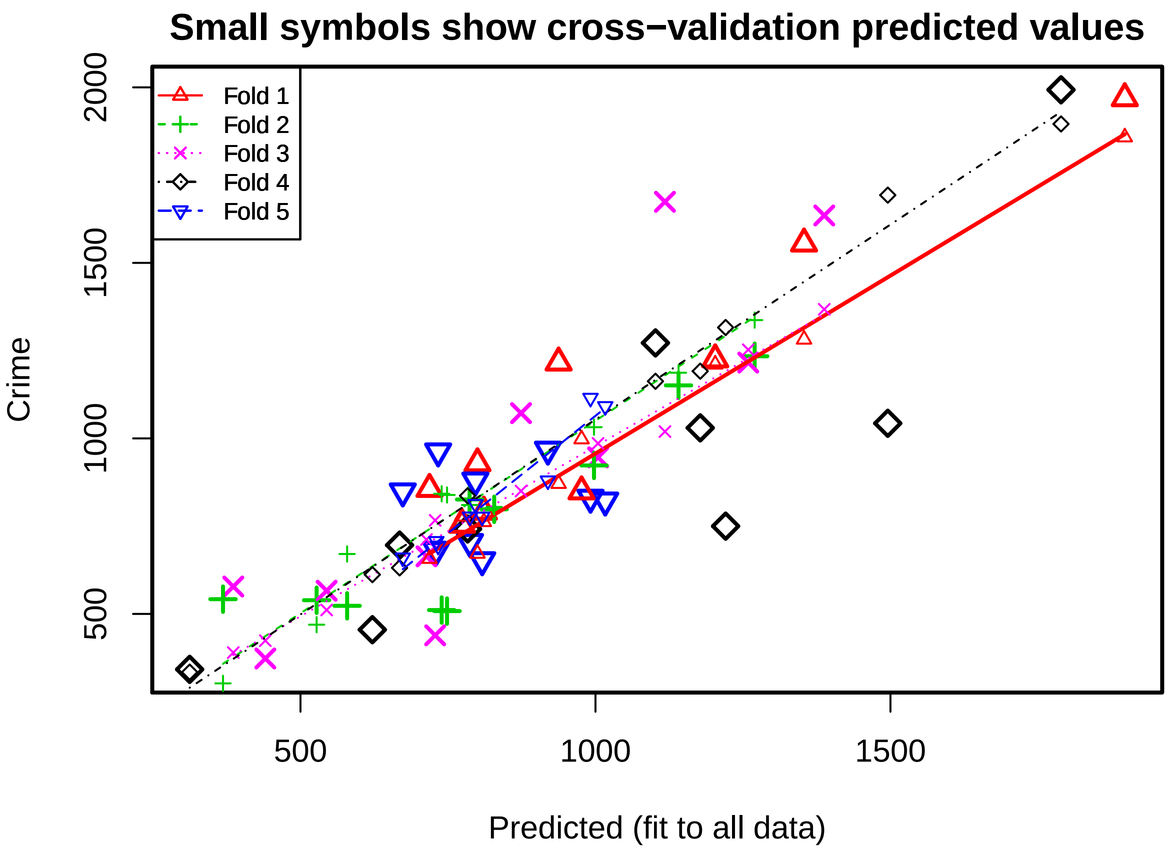
```r
bic2 = BIC(lm.crime2)
bic2
```

```
## [1] 655
```

We'll now perform a cross validation of our improved model using 5 folds.

```r
lm.crime2.cv <- cv.lm(data, lm.crime2, 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 0.000000000053 ***
## U2         1  217386  217386    5.40        0.02534 *
## Ineq       1  848273  848273   21.06 0.000043385425 ***
## 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
```

```r
sse2 = attr(lm.crime2.cv, 'ms') * nrow(data)
sst2 = sum((data$Crime - mean(data$Crime)) ^ 2)
rsquared2 = 1 - sse2/sst2
rsquared2
```

```
## [1] 0.638
```

After performing a K-fold cross validation on our initial and modified models, we find that our R-squared for our improved model is 0.638, a significant improvement over 0.413 We find that the best predictors have changed after performing a K-fold cross validation on our 2nd linear regression model using cv.lm(). Our best predictors are now Ed, Po1, U2, Ineq, and Prob. Resulting in a possibly simpler model. Lets perform a regression now only using significant p-values from this model.

```r
lm.crime3 <-  lm(Crime~Ed+Po1+U2+Ineq+Prob,data=data)
summary(lm.crime3)
```

```
##
## Call:
## lm(formula = Crime ~ Ed + Po1 + U2 + Ineq + Prob, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -562.6 -113.5   14.8  141.5  454.6
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  -3380.3      805.5   -4.20     0.00014 ***
## Ed             168.6       48.4    3.48     0.00120 **
```

```
## Po1            112.0       15.1    7.39 0.0000000046 ***
## U2              44.4       42.3    1.05      0.29980
## Ineq            81.0       14.7    5.53 0.0000020027 ***
## Prob         -3625.1     1685.5   -2.15      0.03743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 222 on 41 degrees of freedom
## Multiple R-squared:  0.708,  Adjusted R-squared:  0.672
## F-statistic: 19.8 on 5 and 41 DF,  p-value: 0.000000000526
```

```
aic3 = AIC(lm.crime3)
aic3
```

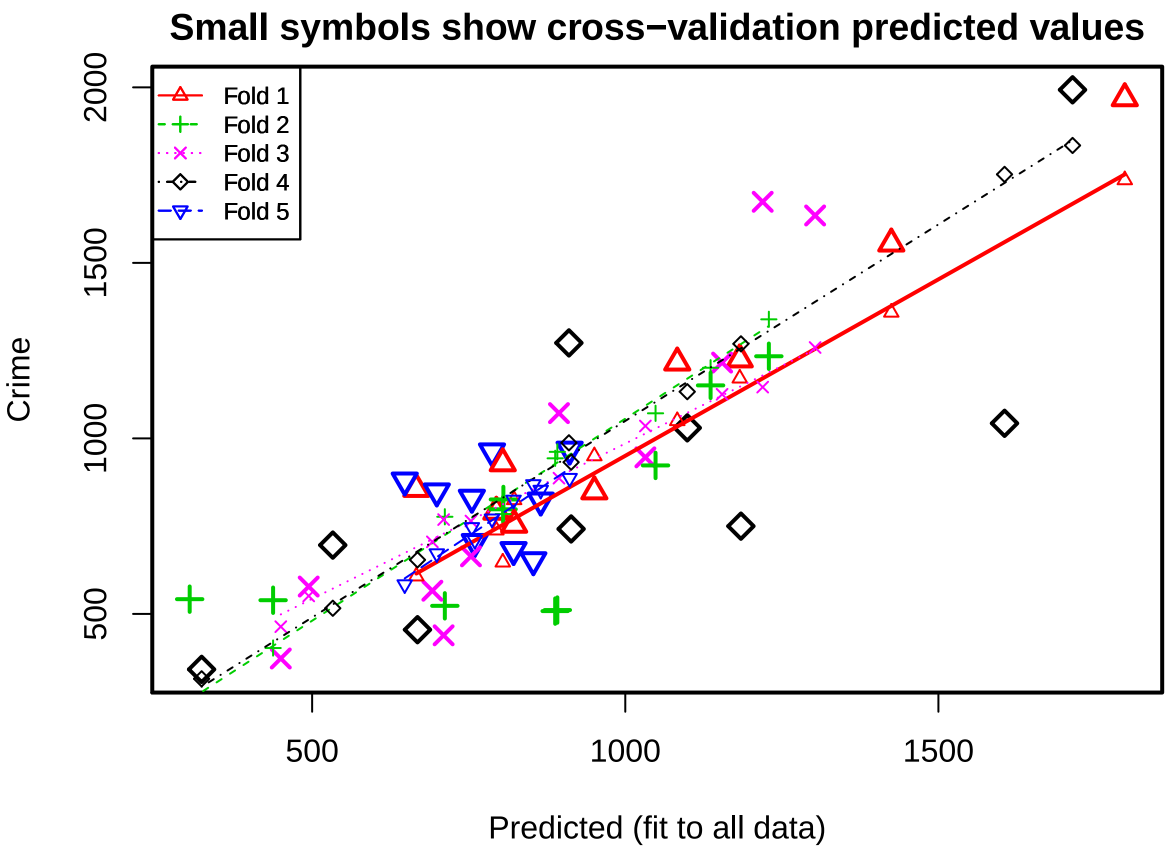
```
## [1] 649
```

```
bic3 = BIC(lm.crime3)
bic3
```

```
## [1] 662
```

```
lm.crime3.cv = cv.lm(data, lm.crime3, m=5)
```

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value       Pr(>F)
## Ed         1  717146  717146   14.62      0.00044 ***
## Po1        1 2536922 2536922   51.71 0.0000000089 ***
## U2         1   17523   17523    0.36      0.55338
## Ineq       1 1370690 1370690   27.94 0.0000044583 ***
## Prob       1  226978  226978    4.63      0.03743 *
## Residuals 41 2011667   49065
## ---
## 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    794.1 1798 1425 666 804 1182.8 1083 822.6  951
## cvpred       738.2 1736 1359 607 647 1171.1 1050 824.6  950
## Crime        791.0 1969 1555 856 929 1225.0 1216 754.0  849
## CV residual   52.8  233  196 249 282   53.9  166 -70.6 -101
##
## Sum of squares = 282666    Mean square = 31407    n = 9
##
## fold 2
## Observations in test set: 10
##                 5  13    15  17   25   34     39     40  42   46
## Predicted    1229 892 804.9 438  712 1048 805.39 1136.2 304  888
## cvpred       1339 962 784.5 403  777 1071 828.63 1201.2 247  943
## Crime        1234 511 798.0 539  523  923 826.00 1151.0 542  508
## CV residual -105 -451  13.5 136 -254 -148  -2.63  -50.2 295 -435
##
## Sum of squares = 598410    Mean square = 59841    n = 10
##
## fold 3
## Observations in test set: 10
##                 2     3   11   14     16   22     28    31   33   38
## Predicted    1303 494.5 1219  754 1032.1  710 1154.6 449.9  894  692
## cvpred       1259 551.2 1146  765 1035.3  769 1125.6 463.7  887  705
## Crime        1635 578.0 1674  664  946.0  439 1216.0 373.0 1072  566
```

```
## CV residual  376  26.8  528 -101  -89.3 -330   90.4 -90.7  185 -139
##
## Sum of squares = 617828    Mean square = 61783    n = 10
##
## fold 4
## Observations in test set: 9
##                19   21   26    27   29  30   36   44   45
## Predicted    1185  914 1714 323.5 1606 533  910 1099  668
## cvpred       1269  932 1834 315.2 1752 516  987 1133  654
## Crime         750  742 1993 342.0 1043 696 1272 1030  455
## CV residual -519 -190  159  26.8 -709 180  285 -103 -199
##
## Sum of squares = 998092    Mean square = 110899    n = 9
##
## fold 5
## Observations in test set: 9
##                 6   7     10  12    24   35    37  41  43
## Predicted     822 787 759.77 699 911.1  853 755.0 648 865
## cvpred        825 771 710.06 672 886.9  869 746.7 584 853
## Crime         682 963 705.00 849 968.0  653 831.0 880 823
## CV residual -143 192  -5.06 177  81.1 -216  84.3 296 -30
##
## Sum of squares = 237446    Mean square = 26383    n = 9
##
## Overall (Sum over all 9 folds)
##    ms
## 58180
```

```
sse3 = attr(lm.crime3.cv, 'ms') * nrow(data)
sst3 = sum((data$Crime - mean(data$Crime)) ^ 2)
rsquared3 = 1 - sse3/sst3
rsquared3
```

```
## [1] 0.603
```

Our modified K-fold validated model only using the significant p-values from the model has the above characteristics. Our model's R-squared has dropped and now only explains approximately 60.261% of our data set's variance. The leaps analysis performed above also confirms that leaving M in as a predictor results in a better model, even though our K-fold validation indicates that M was not significant. Our AIC for all 3 models is 650.029, 640.166, and 648.604. The AIC for all 3 models indicates the 2nd model is the best. The BIC of all 3 models are 681.482, 654.967, and 661.555. The 2nd model of 654.967 is much better than our 1st models BIC of 681.482, and somewhat better than our BIC of the 3rd model of 661.555. The also confirms our 2nd model is likely to be the best. Now that we have our model, let's see how good it is at predicting the point provided in the homework.

```
test_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)

crime_prediction <- predict.lm(lm.crime2, test_point)
crime_prediction
```

```
##    1
## 1304
```

Our final model is then

crime = -5040.5 + (105.0 * M) + (196.5 * Ed) + (115.0 * Po1) + (89.4 * U2) + (67.7 * Ineq) - (3801.8 * Prob)

Our resulting predicted crime rate for the provided point is 1304.245.