STAT425_CaseStudy2

Giang Le and Helen Kim

11/20/2021

```r
# read in the data
uscrime <- read.table("/Users/gianghale/Desktop/fall-2021/stat-425/uscrime.txt", header=T)
dim(uscrime)
```

```
## [1] 47 16
```

```r
summary(uscrime)
```

```
##        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
```

Variable Selection Methods (forward/backward selection using 4 criteria)

```r
# Fit a full model first.
full.model <- lm(Crime ~ ., data=uscrime)
summary(full.model)
```

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

The adjusted R^2 looks quite high but many variables are not statistically signficant. This suggests that there might be multicollinearity and we can select a smaller number of predictors to fit the model.

```
# Using the leaps package to conduct variable selection
library(leaps)
b = regsubsets(Crime ~ ., data=uscrime)
rs = summary(b)
rs$which
```

```
##   (Intercept)     M    So    Ed  Po1   Po2    LF   M.F   Pop    NW    U1    U2
## 1        TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3        TRUE FALSE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4        TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5        TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 6        TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 7        TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 8        TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE
##   Wealth  Ineq  Prob  Time
## 1  FALSE FALSE FALSE FALSE
## 2  FALSE  TRUE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE
```

```
## 4  FALSE  TRUE FALSE FALSE
## 5  FALSE  TRUE  TRUE FALSE
## 6  FALSE  TRUE  TRUE FALSE
## 7   TRUE  TRUE  TRUE FALSE
## 8  FALSE  TRUE  TRUE FALSE
```

Adjusted R^2 as a criteria

```
# Then I examine the R^2 and other criteria such as Cp, AIC, and BIC.
rs$adjr2
```

```
## [1] 0.4610843 0.5612407 0.6423047 0.6718942 0.7059693 0.7307463 0.7341117
## [8] 0.7443692
```

```
which.max(rs$adjr2)
```

```
## [1] 8
```

```
# The best model according to the adjusted R^2 criteria is model 8. The following predictors
# are used in model 8: M, Ed, Po1, M.F, U1, U2, Ineq, Prob
```

Cp as a criteria

```
rs$cp # wants lowest
```

```
## [1] 39.996975 25.070558 13.639362 10.161988  6.257739  3.859603  4.488920
## [8]  4.244947
```

```
which.min(rs$cp)
```

```
## [1] 6
```

```
# The best model according to the Cp-Mallows criteria is model 6. The following predictors
# are used in model 6: M, Ed, Po1, U2, Ineq, Prob.
```

Calculating AIC and BIC for variable selection

```
# I calculated BIC and AIC by hand.
n=dim(uscrime)[1]
msize = 2:9;
BIC = n*log(rs$rss/n) + msize*log(n);
which.min(BIC)
```

```
## [1] 6
```

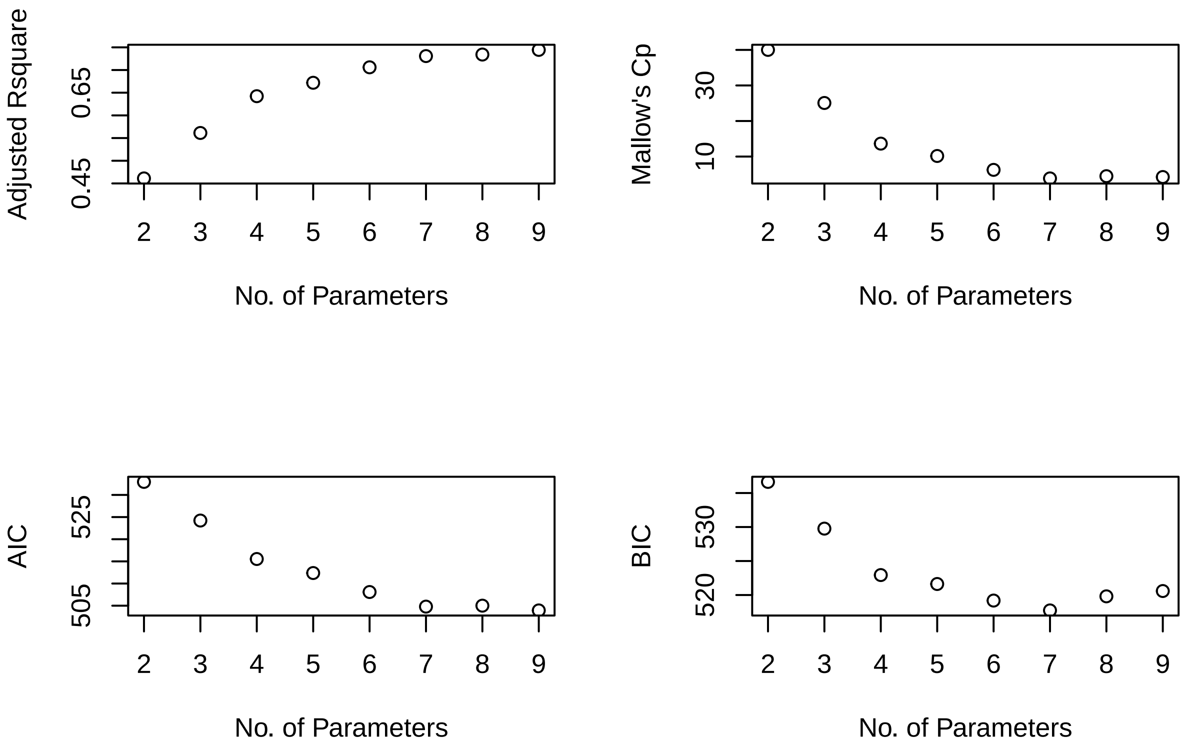
```
AIC = n*log(rs$rss/n) + 2*msize;
which.min(AIC)
```

```
## [1] 8
```

```
# The best model according to the BIC criteria is model 6. The following predictors
# are used in model 6: M, Ed, Po1, U2, Ineq, Prob.
# The best model according to the AIC is model 8. The following predictors
# are used in model 8: M, Ed, Po1, M.F, U1, U2, Ineq, Prob
```

Plotting different criteria

```r
# Verification with plots
par(mfrow=c(2,2))
plot(msize, rs$adjr2, xlab="No. of Parameters", ylab = "Adjusted Rsquare");
plot(msize, rs$cp, xlab="No. of Parameters", ylab = "Mallow's Cp");
plot(msize, AIC, xlab="No. of Parameters", ylab = "AIC");
plot(msize, BIC, xlab="No. of Parameters", ylab = "BIC");
```



Variable selection in both directions

```r
step(full.model, 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 = uscrime)
##
```

```
## 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
```

```
# The best model contains 6 predictors:  Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob.
```

Principal Components Analysis

```
# Check correlation between variables
cor(uscrime[,1:15])
```

```
##                  M          So          Ed         Po1         Po2         LF
## M       1.00000000  0.58435534 -0.53023964 -0.50573690 -0.51317336 -0.1609488
## So      0.58435534  1.00000000 -0.70274132 -0.37263633 -0.37616753 -0.5054695
## Ed     -0.53023964 -0.70274132  1.00000000  0.48295213  0.49940958  0.5611780
## Po1    -0.50573690 -0.37263633  0.48295213  1.00000000  0.99358648  0.1214932
## Po2    -0.51317336 -0.37616753  0.49940958  0.99358648  1.00000000  0.1063496
## LF     -0.16094882 -0.50546948  0.56117795  0.12149320  0.10634960  1.0000000
## M.F    -0.02867993 -0.31473291  0.43691492  0.03376027  0.02284250  0.5135588
## Pop    -0.28063762 -0.04991832 -0.01722740  0.52628358  0.51378940 -0.1236722
## NW      0.59319826  0.76710262 -0.66488190 -0.21370878 -0.21876821 -0.3412144
## U1     -0.22438060 -0.17241931  0.01810345 -0.04369761 -0.05171199 -0.2293997
## U2     -0.24484339  0.07169289 -0.21568155  0.18509304  0.16922422 -0.4207625
## Wealth -0.67005506 -0.63694543  0.73599704  0.78722528  0.79426205  0.2946323
## Ineq    0.63921138  0.73718106 -0.76865789 -0.63050025 -0.64815183 -0.2698865
## Prob    0.36111641  0.53086199 -0.38992286 -0.47324704 -0.47302729 -0.2500861
## Time    0.11451072  0.06681283 -0.25397355  0.10335774  0.07562665 -0.1236404
##                M.F         Pop          NW           U1          U2
## M      -0.02867993 -0.28063762  0.59319826 -0.224380599 -0.24484339
## So     -0.31473291 -0.04991832  0.76710262 -0.172419305  0.07169289
## Ed      0.43691492 -0.01722740 -0.66488190  0.018103454 -0.21568155
## Po1     0.03376027  0.52628358 -0.21370878 -0.043697608  0.18509304
## Po2     0.02284250  0.51378940 -0.21876821 -0.051711989  0.16922422
## LF      0.51355879 -0.12367222 -0.34121444 -0.229399684 -0.42076249
## M.F     1.00000000 -0.41062750 -0.32730454  0.351891900 -0.01869169
## Pop    -0.41062750  1.00000000  0.09515301 -0.038119948  0.27042159
## NW     -0.32730454  0.09515301  1.00000000 -0.156450020  0.08090829
## U1      0.35189190 -0.03811995 -0.15645002  1.000000000  0.74592482
## U2     -0.01869169  0.27042159  0.08090829  0.745924815  1.00000000
## Wealth  0.17960864  0.30826271 -0.59010707  0.044857202  0.09207166
## Ineq   -0.16708869 -0.12629357  0.67731286 -0.063832178  0.01567818
## Prob   -0.05085826 -0.34728906  0.42805915 -0.007469032 -0.06159247
## Time   -0.42769738  0.46421046  0.23039841 -0.169852838  0.10135833
##               Wealth        Ineq         Prob          Time
## M      -0.6700550558  0.63921138  0.361116408  0.1145107190
## So     -0.6369454328  0.73718106  0.530861993  0.0668128312
## Ed      0.7359970363 -0.76865789 -0.389922862 -0.2539735471
## Po1     0.7872252807 -0.63050025 -0.473247036  0.1033577449
## Po2     0.7942620503 -0.64815183 -0.473027293  0.0756266536
## LF      0.2946323090 -0.26988646 -0.250086098 -0.1236404364
## M.F     0.1796086363 -0.16708869 -0.050858258 -0.4276973791
## Pop     0.3082627091 -0.12629357 -0.347289063  0.4642104596
```

```
## NW     -0.5901070652  0.67731286  0.428059153  0.2303984071
## U1      0.0448572017 -0.06383218 -0.007469032 -0.1698528383
## U2      0.0920716601  0.01567818 -0.061592474  0.1013583270
## Wealth  1.0000000000 -0.88399728 -0.555334708  0.0006485587
## Ineq   -0.8839972758  1.00000000  0.465321920  0.1018228182
## Prob   -0.5553347075  0.46532192  1.000000000 -0.4362462614
## Time    0.0006485587  0.10182282 -0.436246261  1.0000000000
```

```
# Some collinearity exists (such as Wealth and Ineq, strongly negative -0.8839972758 or
# Wealth and Ed, Wealth and Po1, Wealth and Po2 etc.
# PCA can help simplifying the data.
```

```
pc <- prcomp(x=uscrime[,1:15], scale=TRUE)
summary(pc)
```

```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377 0.56729
## Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688 0.02145
## Cumulative Proportion  0.4013 0.5880 0.7217 0.79920 0.86308 0.89996 0.92142
##                            PC8     PC9    PC10    PC11    PC12    PC13   PC14
## Standard deviation     0.55444 0.48493 0.44708 0.41915 0.35804 0.26333 0.2418
## Proportion of Variance 0.02049 0.01568 0.01333 0.01171 0.00855 0.00462 0.0039
## Cumulative Proportion  0.94191 0.95759 0.97091 0.98263 0.99117 0.99579 0.9997
##                           PC15
## Standard deviation     0.06793
## Proportion of Variance 0.00031
## Cumulative Proportion  1.00000
```

```
# Using to this analysis,  using the first 6 principal components can help explain
# 90% variance in the data.
```

```
# Rerunnning regression using only the first 6 components from PCA.
modpcr<-lm(uscrime[,16] ~ pc$x[,1:6])
summary(modpcr)
```

```
##
## Call:
## lm(formula = uscrime[, 16] ~ pc$x[, 1:6])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -377.15 -172.23   25.81  132.10  480.38
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      905.09      35.35  25.604  < 2e-16 ***
## pc$x[, 1:6]PC1    65.22      14.56   4.478 6.14e-05 ***
## pc$x[, 1:6]PC2   -70.08      21.35  -3.283  0.00214 **
## pc$x[, 1:6]PC3    25.19      25.23   0.998  0.32409
## pc$x[, 1:6]PC4    69.45      33.14   2.095  0.04252 *
## pc$x[, 1:6]PC5  -229.04      36.50  -6.275 1.94e-07 ***
## pc$x[, 1:6]PC6   -60.21      48.04  -1.253  0.21734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 242.3 on 40 degrees of freedom
## Multiple R-squared:  0.6586, Adjusted R-squared:  0.6074
## F-statistic: 12.86 on 6 and 40 DF,  p-value: 4.869e-08
```

```
# Even though R^2 decreased compared to the original full model but the number
# of stat. significant predictors increased. It is likely that there is
# a reduction in the test errors compared to the full model. Let's check this
# by dividing the dataset into a train and test set and evaluate the training
# and test errors across models.
```

```
set.seed(425)
ntrain <- round(n*0.7) # use 70% of the data for training
tindex <- sample(n, ntrain)
train <- uscrime[tindex,]
test <- uscrime[-tindex,]
dim(train)
```

Evaluating PCA

```
## [1] 33 16
```

```
dim(test)
```

```
## [1] 14 16
```

```
# Now I fit a linear model using only the train data only.
lm_model <- lm(Crime ~ ., data=train)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -344.57  -94.11    4.73  100.75  385.52
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.067e+03  2.013e+03  -3.511  0.00268 **
## M            3.514e+01  5.850e+01   0.601  0.55595
## So           9.777e+01  2.274e+02   0.430  0.67260
## Ed           2.282e+02  7.656e+01   2.981  0.00839 **
## Po1          1.070e+02  1.475e+02   0.725  0.47809
## Po2         -8.608e+00  1.589e+02  -0.054  0.95744
## LF          -3.707e+02  1.947e+03  -0.190  0.85124
## M.F          3.729e+01  2.791e+01   1.336  0.19914
## Pop          4.648e-01  1.689e+00   0.275  0.78649
## NW           4.748e+00  8.166e+00   0.581  0.56856
## U1          -5.282e+03  5.518e+03  -0.957  0.35188
## U2           1.634e+02  1.222e+02   1.337  0.19878
## Wealth      -8.097e-02  1.595e-01  -0.508  0.61816
## Ineq         3.793e+01  4.186e+01   0.906  0.37763
## Prob        -8.465e+02  2.760e+03  -0.307  0.76280
## Time         1.394e+01  1.008e+01   1.384  0.18430
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 216.2 on 17 degrees of freedom
## Multiple R-squared:  0.8422, Adjusted R-squared:  0.7029
## F-statistic: 6.048 on 15 and 17 DF,  p-value: 0.0003347
```

```
# Root mean squared error of training data. (155.1958)
rmse<-function(x,y) sqrt(mean((x-y)^2))
rmse(fitted(lm_model), train$Crime)
```

```
## [1] 155.1958
```

```
# Root mean squared error of testing data. (286.0712)
rmse(predict(lm_model,test), test$Crime)
```

```
## [1] 286.0712
```

```
# Compared with the PCA model modpcr
rmse(fitted(modpcr), train$Crime)
```

```
## Warning in x - y: longer object length is not a multiple of shorter object
## length

## [1] 444.1344
```

```
rmse(predict(modpcr,test), test$Crime)
```

```
## Warning: 'newdata' had 14 rows but variables found have 47 rows

## Warning: longer object length is not a multiple of shorter object length

## [1] 519.5657
```

```
# We can see that the RMSE for the test data is larger than that of the train data.
# in both cases. This might be due to the small number of observations available.
```

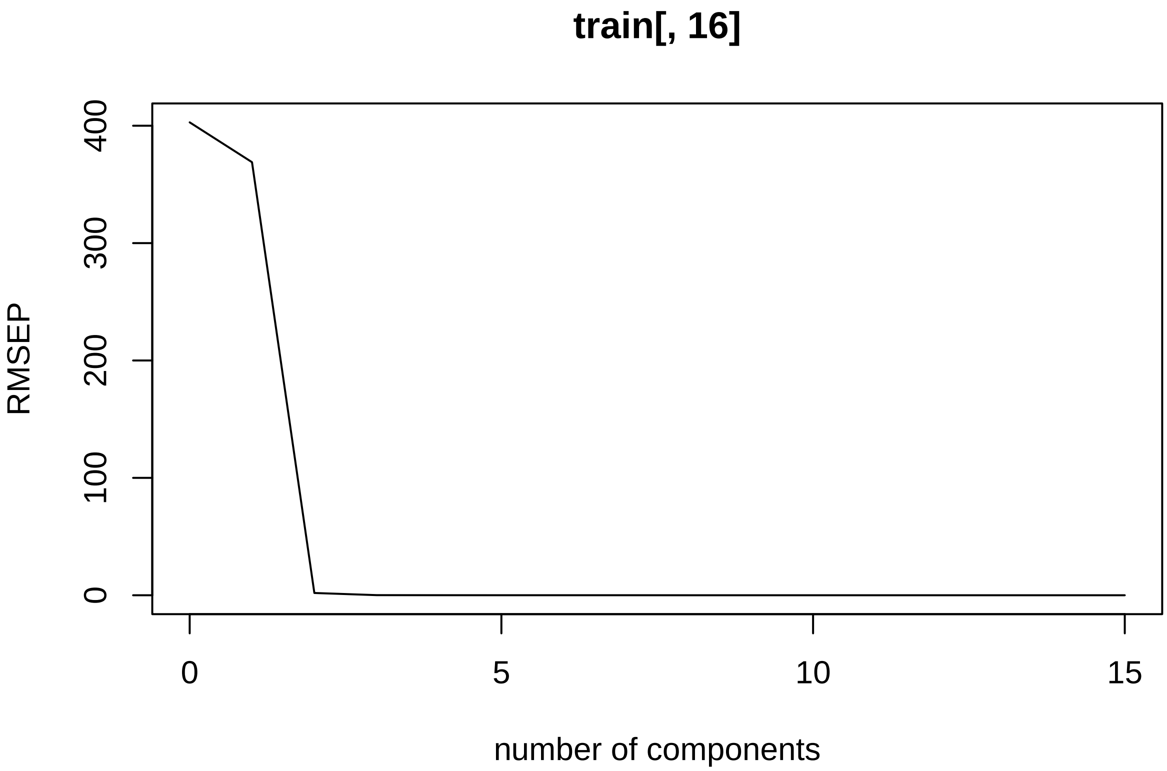
Let's try to use CV to determine the number of PC needed.

```
library(pls)
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
set.seed(425)
modpcrcv<-pcr(train[,16] ~.,data=train,validation="CV",ncomp=15)
pcrCV<-RMSEP(modpcrcv,estimate="CV")
plot(pcrCV)
```



According to the plot above we could get away with 2 components?

```
pcpred<-predict(modpcrcv,test,ncomp=2)
rmse(pcpred,test$Crime) # We have very small RMSE. Let's try to fit a linear regression model with two
```

```
## [1] 1.650941
```

```
modpcr1<-lm(uscrime[,16] ~ pc$x[,1:2])
summary(modpcr1)
```

```
##
## Call:
## lm(formula = uscrime[, 16] ~ pc$x[, 1:2])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -575.1 -222.1    4.4  159.6  905.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       905.09      49.52  18.279  < 2e-16 ***
## pc$x[, 1:2]PC1    65.22      20.40   3.197  0.00257 **
## pc$x[, 1:2]PC2   -70.08      29.90  -2.344  0.02367 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 339.5 on 44 degrees of freedom
## Multiple R-squared:  0.2631, Adjusted R-squared:  0.2296
## F-statistic: 7.856 on 2 and 44 DF,  p-value: 0.001209
```

Use Ridge Regression