HW8.2

Question 8.2:

Using crime data from http://www.statsci.org/data/general/uscrime.txt (http://www.statsci.org/data/general/uscrime.txt) (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html (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:

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

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Steps:

1. Set up libraries and data
2. Determine what model to use - factors, any transformations, etc.
    a. Build lm model
    b. Use adjusted R-squared, BIC, AIC, and MSE to evaluate model quality. Check normality assumptions.
    c. Remove variable with least significant p-value
    d. Repeat these steps until removing more variables decreases model quality
    e. Check effect of adding interaction terms
3. Use final model to predict crime rate for test datapoint above

Step 1: Set up libraries and data

```
# Import libraries
library(ggplot2)
library(tidyverse)
library(modelr)
library(broom)
library(GGally)

# Read data
crime <- read.delim("uscrime.txt", header = T)
head(crime)
```

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

Step 2: Determine what model to use

Let's start with a linear regression model using all 15 predictor variables.

```
m <- lm(crime$Crime~., data=crime)
summary(m)
```

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

And check model quality:

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

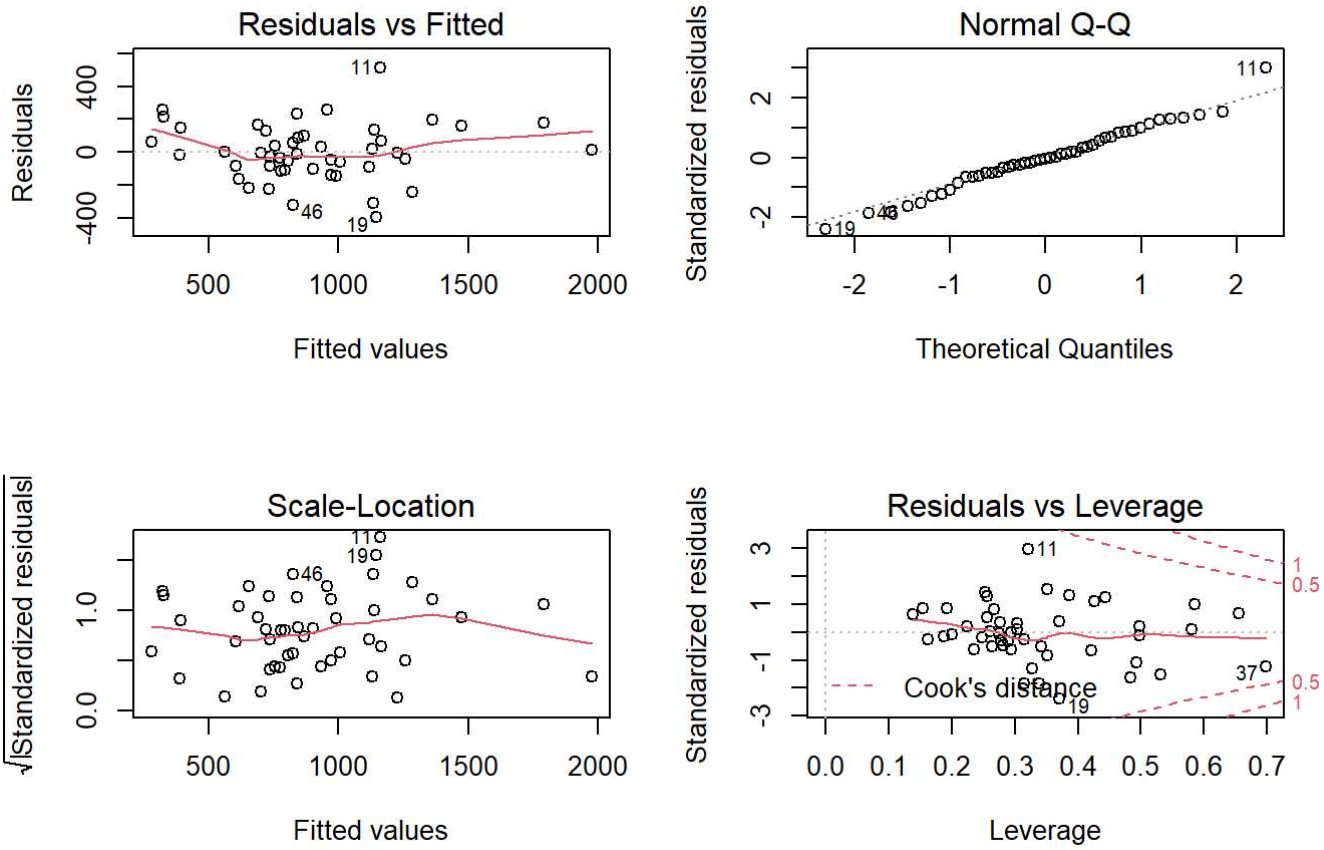
```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.708  650.  681.
```

```
paste("MSE", mean(m$residuals^2))
```

```
## [1] "MSE 28828.6334305033"
```

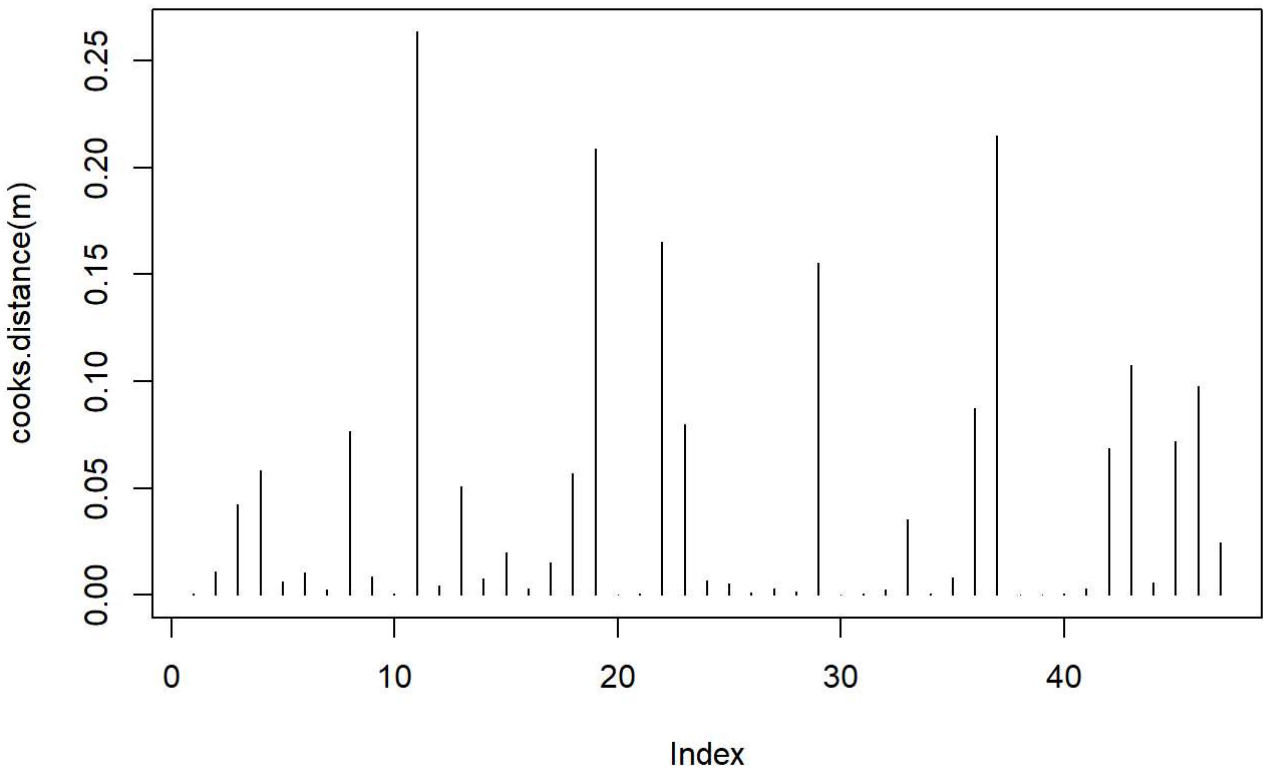
We can see that only a few of the predictor variables are currently statistically significant. Even so, this model's adjusted r-squared is 70%, meaning it can explain 70% of the variation in the value of the response variable. I am using adjusted R-squared, not R-squared, because having a lot of predictors can inflate R-squared but adjusted R-squared compensates for this. Let's check model normality assumptions:

```
par(mfrow = c(2,2))
plot(m)
```



Residuals vs. fitted plot looks reasonable - no clear trend. It doesn't look like we need to do any transformations on the data. Residuals are generally normally distributed, although point 11 does stand out. Let's check a clearer Cook's distance plot:

```
plot(cooks.distance(m), type="h")
```



Points 11, 19, and 37 may be more influential in the model. However, they don't exceed 1.

Let's start by removing the variable with the least significant p-value, So, rerun the model, and compare.

```
# Remove So
m <- lm(Crime~M+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW +
##     U1 + U2 + Wealth + Ineq + Prob + Time, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.72  -98.25   -6.12  112.90  513.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.981e+03  1.597e+03  -3.746 0.000711 ***
## M            8.773e+01  4.087e+01   2.146 0.039520 *
## Ed           1.882e+02  6.100e+01   3.086 0.004170 **
## Po1          1.927e+02  1.044e+02   1.846 0.074152 .
## Po2         -1.092e+02  1.154e+02  -0.947 0.350855
## LF          -6.461e+02  1.275e+03  -0.507 0.615736
## M.F          1.733e+01  1.979e+01   0.875 0.387876
## Pop         -7.331e-01  1.269e+00  -0.578 0.567573
## NW           4.135e+00  5.786e+00   0.715 0.480002
## U1          -5.786e+03  3.834e+03  -1.509 0.141106
## U2           1.673e+02  7.902e+01   2.118 0.042081 *
## Wealth       9.555e-02  9.928e-02   0.962 0.343041
## Ineq         7.045e+01  2.074e+01   3.398 0.001834 **
## Prob        -4.864e+03  2.213e+03  -2.197 0.035344 *
## Time        -3.455e+00  6.991e+00  -0.494 0.624556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205.8 on 32 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7169
## F-statistic: 9.322 on 14 and 32 DF,  p-value: 1.118e-07
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.717  648.  678.
```

Adjusted R-squared has increased, and AIC and BIC have both decreased, so the model has improved.

Let's continue using p-values to remove variables 1 by 1 and inspecting the results:

Remove Time:

```
# Remove Time
m <- lm(Crime~M+Ed+Po1+Po2+LF+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW +
##     U1 + U2 + Wealth + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -385.12  -98.09    6.99  108.57  486.83
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.285e+03  1.456e+03  -4.318 0.000135 ***
## M            8.475e+01  3.996e+01   2.121 0.041529 *
## Ed           1.895e+02  6.024e+01   3.146 0.003497 **
## Po1          1.772e+02  9.842e+01   1.801 0.080914 .
## Po2         -8.943e+01  1.069e+02  -0.836 0.409032
## LF          -6.357e+02  1.260e+03  -0.505 0.617215
## M.F          1.927e+01  1.917e+01   1.005 0.322234
## Pop         -8.847e-01  1.217e+00  -0.727 0.472563
## NW           3.371e+00  5.511e+00   0.612 0.544915
## U1          -5.610e+03  3.774e+03  -1.486 0.146643
## U2           1.643e+02  7.788e+01   2.110 0.042553 *
## Wealth       9.129e-02  9.777e-02   0.934 0.357237
## Ineq         7.125e+01  2.044e+01   3.486 0.001406 **
## Prob        -4.213e+03  1.758e+03  -2.396 0.022409 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.4 on 33 degrees of freedom
## Multiple R-squared:  0.8016, Adjusted R-squared:  0.7234
## F-statistic: 10.25 on 13 and 33 DF,  p-value: 3.761e-08
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.723  646.  674.
```

Remove LF:

```
# Remove LF
m <- lm(Crime~M+Ed+Po1+Po2+M.F+Pop+NW+U1+U2+Wealth+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -390.49 -100.89    6.44  109.59  482.43
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.199e+03  1.430e+03  -4.336 0.000122 ***
## M            8.930e+01  3.850e+01   2.320 0.026506 *
## Ed           1.793e+02  5.610e+01   3.196 0.003008 **
## Po1          1.655e+02  9.461e+01   1.750 0.089178 .
## Po2         -7.353e+01  1.011e+02  -0.728 0.471886
## M.F          1.436e+01  1.634e+01   0.879 0.385674
## Pop         -9.847e-01  1.188e+00  -0.829 0.412947
## NW           2.882e+00  5.365e+00   0.537 0.594679
## U1          -5.013e+03  3.544e+03  -1.414 0.166336
## U2           1.669e+02  7.685e+01   2.171 0.036996 *
## Wealth       8.993e-02  9.665e-02   0.930 0.358685
## Ineq         7.110e+01  2.021e+01   3.518 0.001256 **
## Prob        -4.020e+03  1.697e+03  -2.368 0.023696 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201.2 on 34 degrees of freedom
## Multiple R-squared:    0.8,  Adjusted R-squared:  0.7295
## F-statistic: 11.34 on 12 and 34 DF,  p-value: 1.209e-08
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.729  645.  671.
```

Remove NW:

```
# Remove NW
m <- lm(Crime~M+Ed+Po1+Po2+M.F+Pop+U1+U2+Wealth+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 +
##     Wealth + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -406.59 -102.72   10.15  118.72  492.79
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.223e+03  1.414e+03  -4.400 9.68e-05 ***
## M            9.756e+01  3.493e+01   2.793 0.008413 **
## Ed           1.758e+02  5.516e+01   3.187 0.003019 **
## Po1          1.612e+02  9.330e+01   1.728 0.092838 .
## Po2         -6.391e+01  9.845e+01  -0.649 0.520470
## M.F          1.317e+01  1.602e+01   0.822 0.416654
## Pop         -9.468e-01  1.174e+00  -0.807 0.425341
## U1          -5.097e+03  3.504e+03  -1.455 0.154689
## U2           1.707e+02  7.573e+01   2.254 0.030544 *
## Wealth       8.495e-02  9.522e-02   0.892 0.378427
## Ineq         7.411e+01  1.922e+01   3.857 0.000471 ***
## Prob        -3.743e+03  1.601e+03  -2.338 0.025211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 199.1 on 35 degrees of freedom
## Multiple R-squared:  0.7984, Adjusted R-squared:  0.735
## F-statistic:  12.6 on 11 and 35 DF,  p-value: 3.746e-09
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.735  643.  667.
```

Remove Po2:

```
# Remove Po2
m <- lm(Crime~M+Ed+Po1+M.F+Pop+U1+U2+Wealth+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth +
##     Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -441.66 -103.21   19.53  122.96  474.69
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.404e+03  1.375e+03  -4.656 4.28e-05 ***
## M            9.720e+01  3.465e+01   2.805 0.008054 **
## Ed           1.718e+02  5.437e+01   3.160 0.003192 **
## Po1          1.020e+02  1.966e+01   5.189 8.43e-06 ***
## M.F          1.515e+01  1.560e+01   0.971 0.337823
## Pop         -8.776e-01  1.159e+00  -0.757 0.454048
## U1          -5.170e+03  3.474e+03  -1.488 0.145442
## U2           1.725e+02  7.508e+01   2.297 0.027539 *
## Wealth       8.574e-02  9.445e-02   0.908 0.370044
## Ineq         7.507e+01  1.900e+01   3.950 0.000349 ***
## Prob        -3.759e+03  1.588e+03  -2.367 0.023411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 197.5 on 36 degrees of freedom
## Multiple R-squared:  0.7959, Adjusted R-squared:  0.7392
## F-statistic: 14.04 on 10 and 36 DF,  p-value: 1.169e-09
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.739  642.  664.
```

Remove Wealth:

```
# Remove Wealth
m <- lm(Crime~M+Ed+Po1+M.F+Pop+U1+U2+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Ineq +
##     Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.97 -104.20    8.31  117.37  502.63
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6068.1757  1321.6242  -4.591 4.95e-05 ***
## M              91.1349    33.9155   2.687  0.01073 *
## Ed            181.4908    53.1944   3.412  0.00158 **
## Po1           108.7265    18.1775   5.981 6.65e-07 ***
## M.F            17.9369    15.2604   1.175  0.24735
## Pop            -0.7533     1.1486  -0.656  0.51600
## U1          -5855.3447  3383.0241  -1.731  0.09182 .
## U2            187.5617    73.0352   2.568  0.01440 *
## Ineq           64.2130    14.7338   4.358  0.00010 ***
## Prob        -4050.5933  1551.2988  -2.611  0.01296 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 197 on 37 degrees of freedom
## Multiple R-squared:  0.7913, Adjusted R-squared:  0.7405
## F-statistic: 15.58 on 9 and 37 DF,  p-value: 4.169e-10
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.740  641.  661.
```

Remove Pop:

```
# Remove Pop
m <- lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.70 -111.07    3.03  122.15  483.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6426.10    1194.61  -5.379 4.04e-06 ***
## M              93.32      33.50   2.786  0.00828 **
## Ed            180.12      52.75   3.414  0.00153 **
## Po1           102.65      15.52   6.613 8.26e-08 ***
## M.F            22.34      13.60   1.642  0.10874
## U1          -6086.63    3339.27  -1.823  0.07622 .
## U2            187.35      72.48   2.585  0.01371 *
## Ineq           61.33      13.96   4.394 8.63e-05 ***
## Prob        -3796.03    1490.65  -2.547  0.01505 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195.5 on 38 degrees of freedom
## Multiple R-squared:  0.7888, Adjusted R-squared:  0.7444
## F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.744  639.  658.
```

Remove M.F:

```
# Remove M.F
m <- lm(Crime~M+Ed+Po1+U1+U2+Ineq+Prob, data=crime)
summary(m)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U1 + U2 + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -520.76 -105.67    9.53  136.28  519.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5095.55     896.90  -5.681 1.43e-06 ***
## M             106.78      33.18   3.218   0.0026 **
## Ed            218.45      48.33   4.520 5.62e-05 ***
## Po1           105.96      15.72   6.738 4.91e-08 ***
## U1          -3542.35    3021.94  -1.172   0.2482
## U2            158.82      71.89   2.209   0.0331 *
## Ineq           66.33      13.92   4.767 2.61e-05 ***
## Prob        -3730.85    1522.21  -2.451   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 199.8 on 39 degrees of freedom
## Multiple R-squared:  0.7738, Adjusted R-squared:  0.7332
## F-statistic: 19.06 on 7 and 39 DF,  p-value: 8.805e-11
```

```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.733  641.  657.
```

This model is a little worse than the previous one. R-squared has decreased and AIC has increased slightly, but BIC has decreased slightly. Let's try removing one more variable and see if the model improves or continues to get worse.

Remove U1:

```
# Remove U1
m <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob, data=crime)
summary(m)
```

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

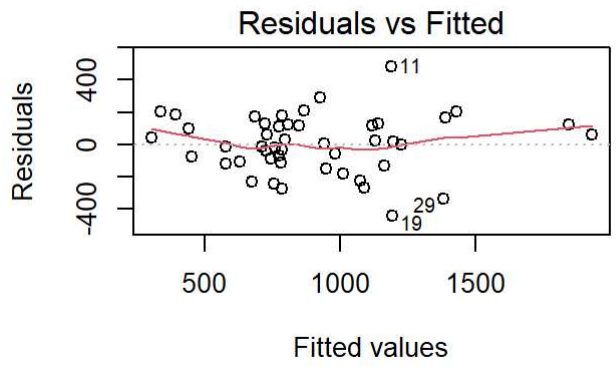
```
glance(m)[,c("adj.r.squared", "AIC", "BIC")]
```

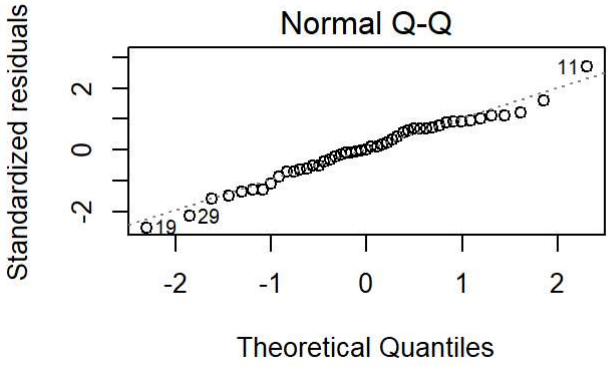
```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.731  640.  655.
```

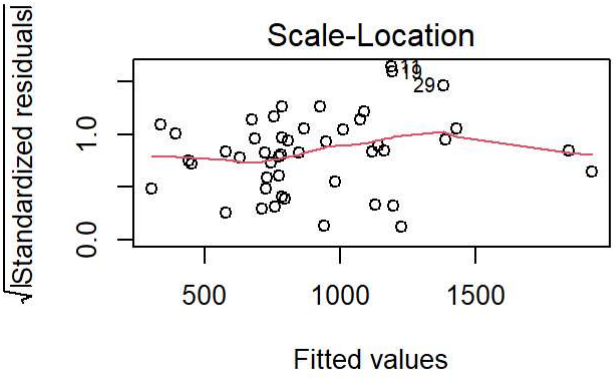
Model has not improved based on adjusted R-squared and AIC. Let's add U1 and M.F back in as the model performed better before we removed them.

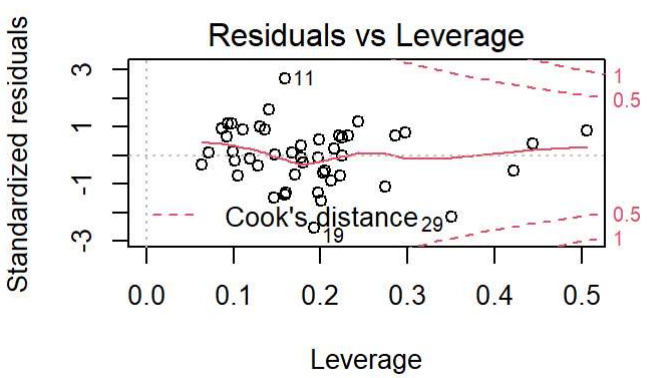
We will use M, Ed, Po1, M.F, U1, U2, Ineq, and Prob as explanatory variables since this model performed best overall. Let's take a look at this model's plots to check for normality:

```
mbest <- lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob, data=crime)
par(mfrow=c(2,2))
plot(mbest)
```







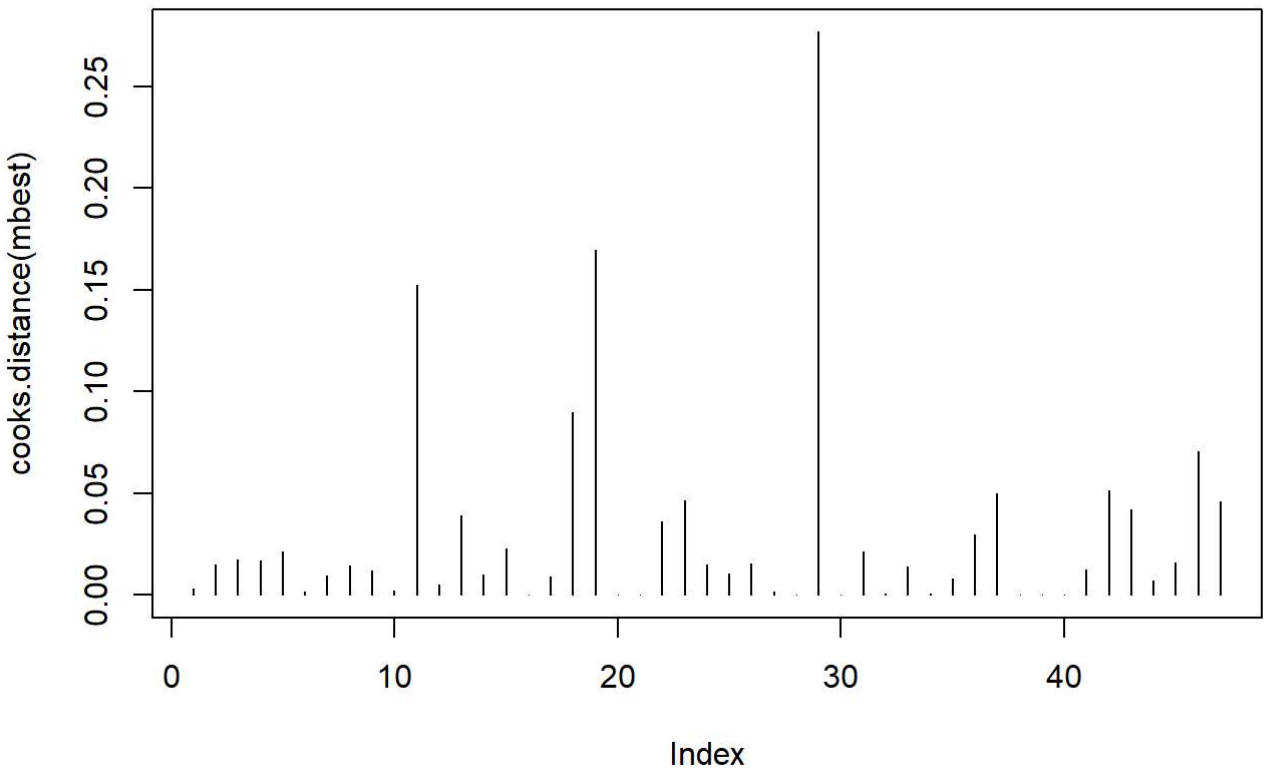



```
shapiro.test(mbest$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mbest$residuals
## W = 0.98511, p-value = 0.8051
```

Residuals pass Shapiro-Wilk test for normality, so we can assume normal distribution.

```
plot(cooks.distance(mbest), type="h")
```



Overall, this model looks reasonable with regards to normality assumptions - there isn't a clear pattern in the fitted vs. residual plot, and residuals are more normally distributed than in our original model. Cook's distance also shows fewer points with a high influence (point 29 stands out the most at .25).

Let's get VIF to see if we need to worry about multicollinearity:

```
car::vif(mbest)
```

```
##        M       Ed      Po1      M.F       U1       U2     Ineq     Prob
## 2.131963 4.189684 2.560496 1.932367 4.360038 4.508106 3.731074 1.381879
```

Generally, VIF values above 5 or 10 are cause for concern. We don't have any values quite that high in this model, but Ed, U1, and U2 are close to 5.

R also has the step() function in the stats library, which uses backwards selection to pick the best model, based on optimizing for AIC. Let's compare that to the model we got by manually eliminating factors:

```
m <- lm(Crime~., data=crime)
best_step_m <- step(m)
```

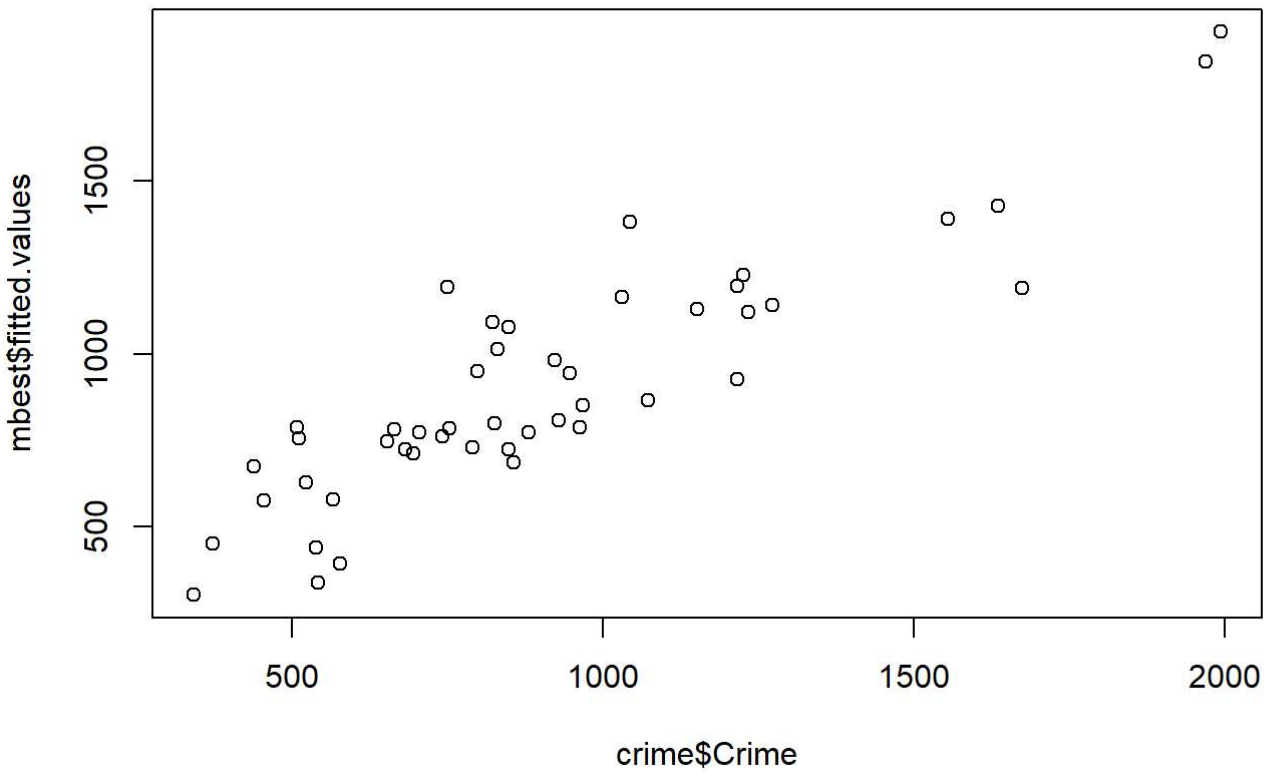
```
## Start:  AIC=514.65
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - So      1        29 1354974 512.65
## - LF      1      8917 1363862 512.96
## - Time    1     10304 1365250 513.00
## - Pop     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
## - 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
## - 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
## - 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
## - 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
## - 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
## - 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
## - M.F   1    103159 1556227 505.16
## - 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
```

Interestingly, step() selected the exact same model.

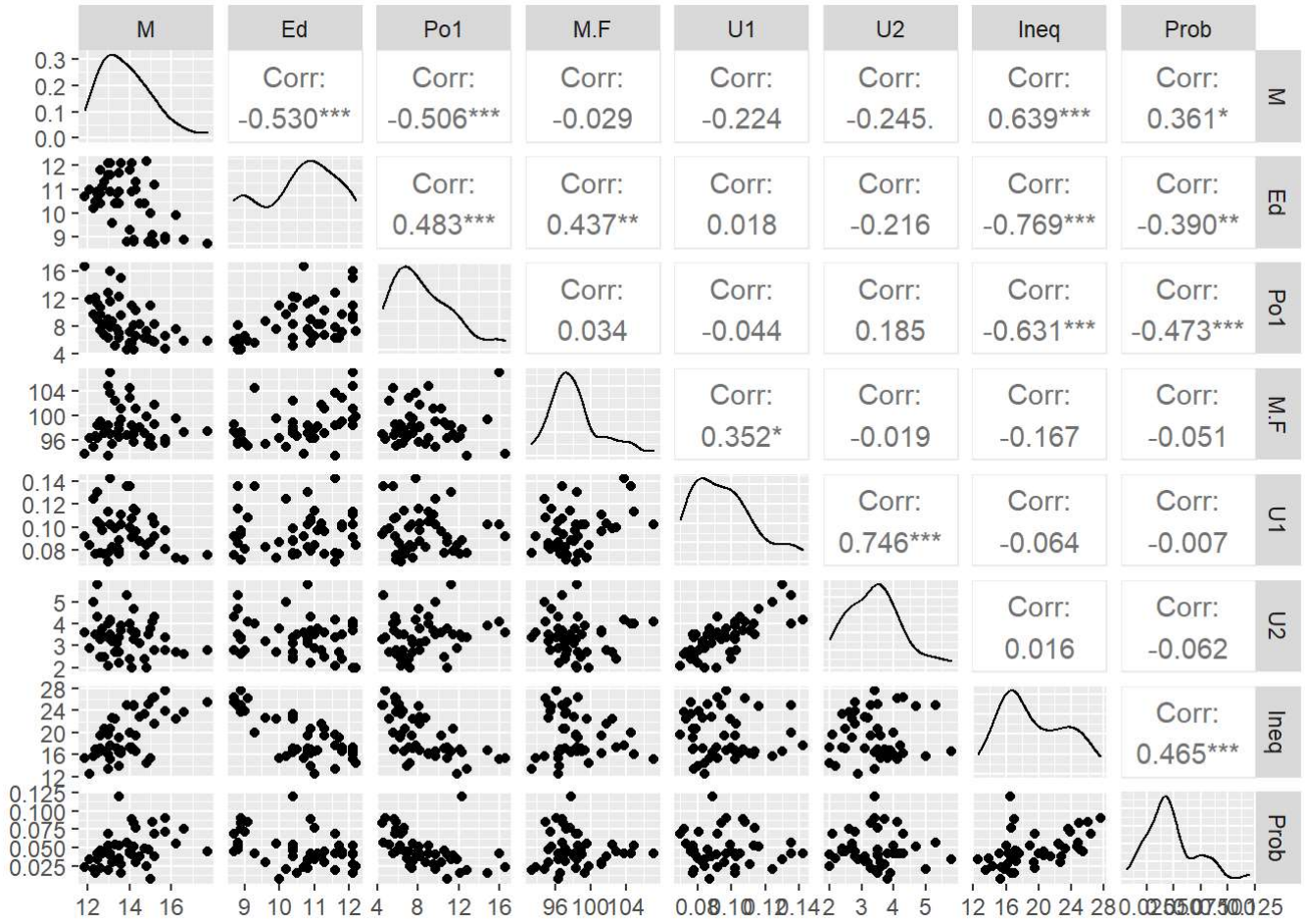
```
# Plot fitted vs. actual values
plot(x=crime$Crime, y=mbest$fitted.values)
```



There is a fairly clear linear relationship between predicted and actual values for crime using this model.

Next, we can look into adding interaction terms to this model, to see if we can further increase adjusted R-squared while decreasing AIC and BIC. Interaction terms can help account for the interactions between variables in the model. I am looking forward to learning more about variable selection in class, but in the meantime, here is my thinking: try to account for the interaction between variables that are more closely correlated. Looking at the correlation plot below, I tried adding in interaction terms for pairs of variables that had a stronger correlation, and kept the ones that improved the model's quality. You can see the successful ones below:

```
cols <- c("M", "Ed", "Po1", "M.F", "U1", "U2", "Ineq", "Prob")
ggpairs(crime[cols])
```




```
# Add U1*U2
glance(lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1:U2, data=crime))[,c("adj.r.squared", "AIC", "BI
C")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.746  640.  660.
```

```
# Add Ed*Ineq
glance(lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1:U2+Ed:Ineq, data=crime))[,c("adj.r.squared", "AI
C", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.758  638.  660.
```

```
# Add M*Ineq
glance(lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1:U2+Ed:Ineq+M:Ineq, data=crime))[,c("adj.r.square
d", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.786  633.  657.
```

```
# Add Po1*Ineq
glance(lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1:U2+Ed:Ineq+M:Ineq+Po1:Ineq, data=crime))[,c("ad
j.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.796  632.  657.
```

Let's compare the best models with and without interaction terms:

```
# Compare model with interaction terms to model without:
minter <- lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1:U2+Ed:Ineq+M:Ineq+Po1:Ineq+Ed:Po1, data=crim
e)
# Model with interaction terms:
glance(minter)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.802  631.  658.
```

```
# Model without interaction terms:
glance(mbest)[,c("adj.r.squared", "AIC", "BIC")]
```

```
## # A tibble: 1 x 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.744  639.  658.
```

```
paste("MSE no interaction terms", mean(mbest$residuals^2))
```

```
## [1] "MSE no interaction terms 30916.3354924821"
```

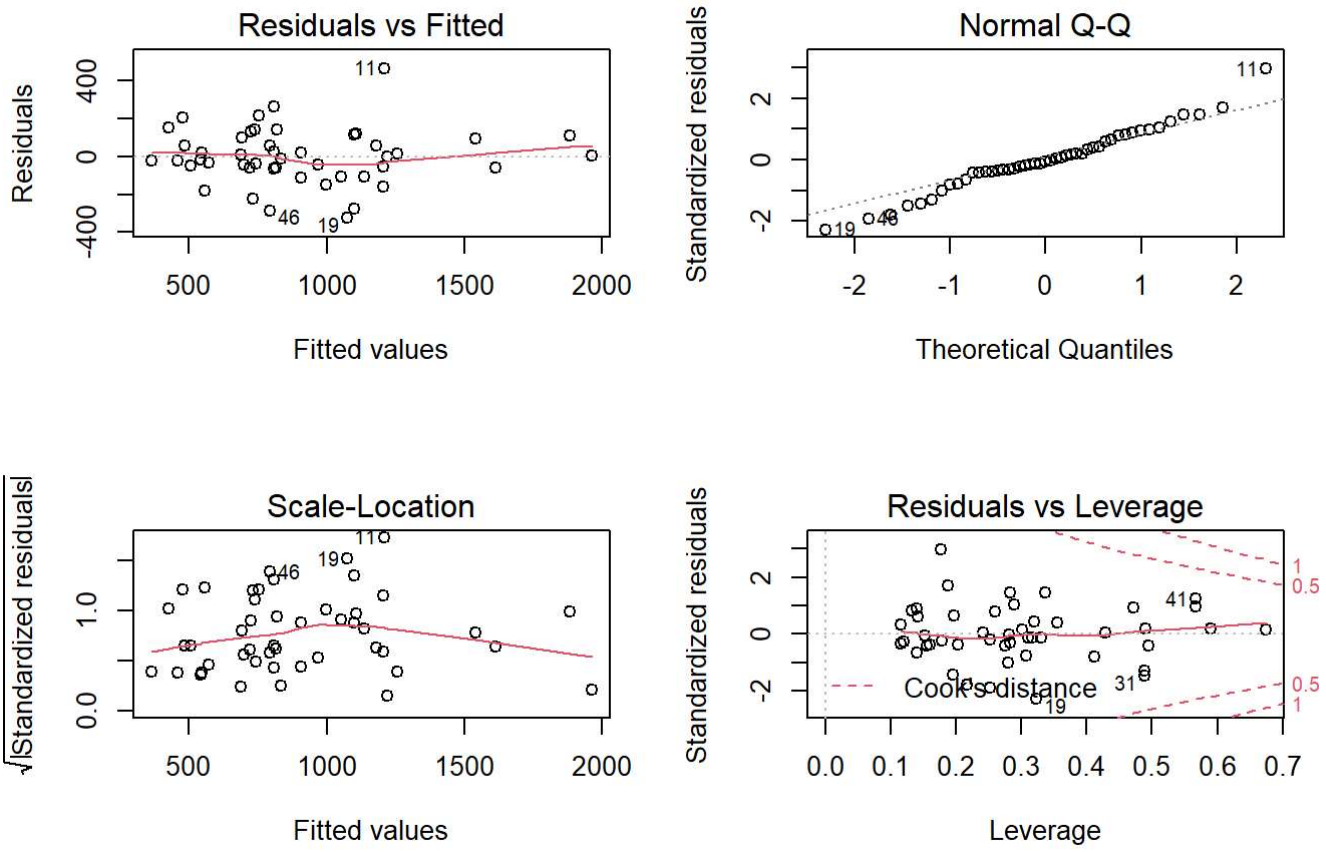
```
paste("MSE interaction terms", mean(minter$residuals^2))
```

```
## [1] "MSE interaction terms 20816.7234133646"
```

This model with added interaction terms has a higher adjusted R-squared and lower AIC. MSE was significantly lower for the model with interaction terms. BIC is the same. Overall, the model with interaction terms performed better on 3 out of 4 measures, and the same on 1. Let's check the normality assumptions for the model with

interaction terms:

```
par(mfrow=c(2,2))
plot(minter)
```



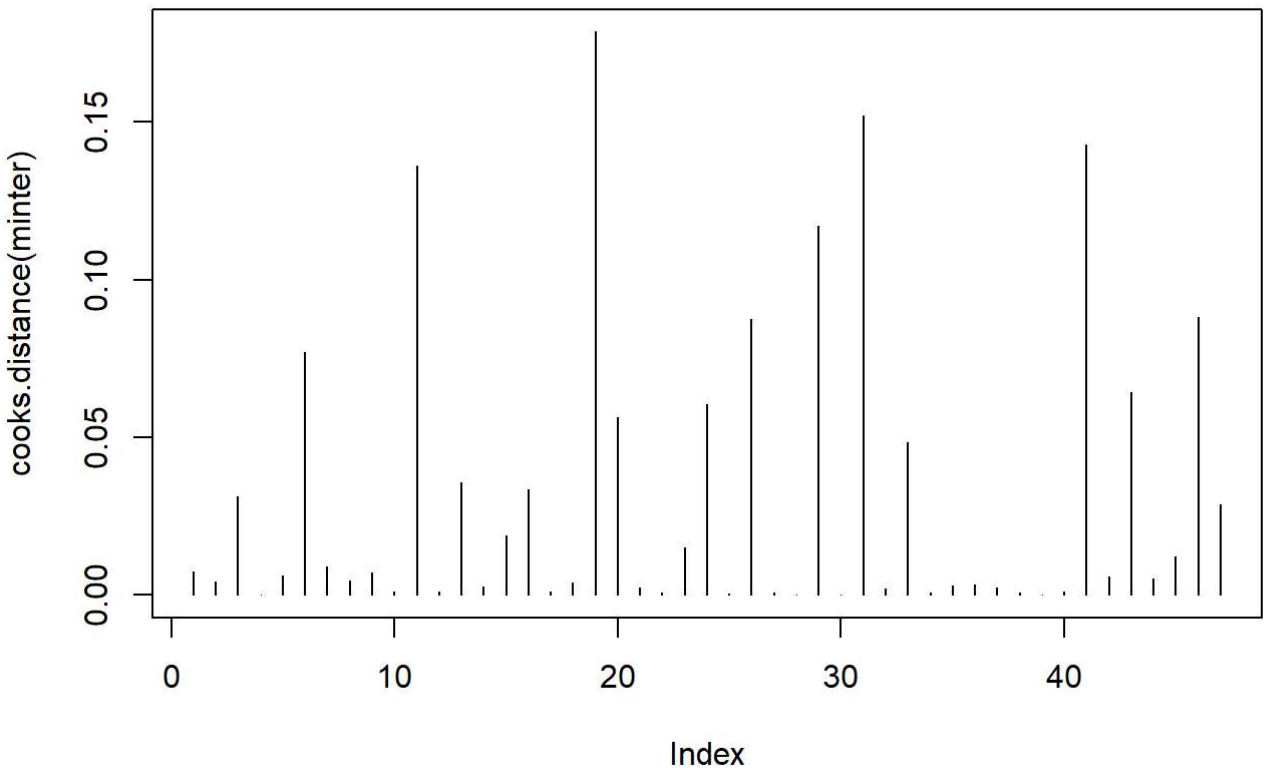
Residuals vs. fitted plot has further improved, with trend close to being a flat line at 0.

```
shapiro.test(minter$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  minter$residuals
## W = 0.9697, p-value = 0.258
```

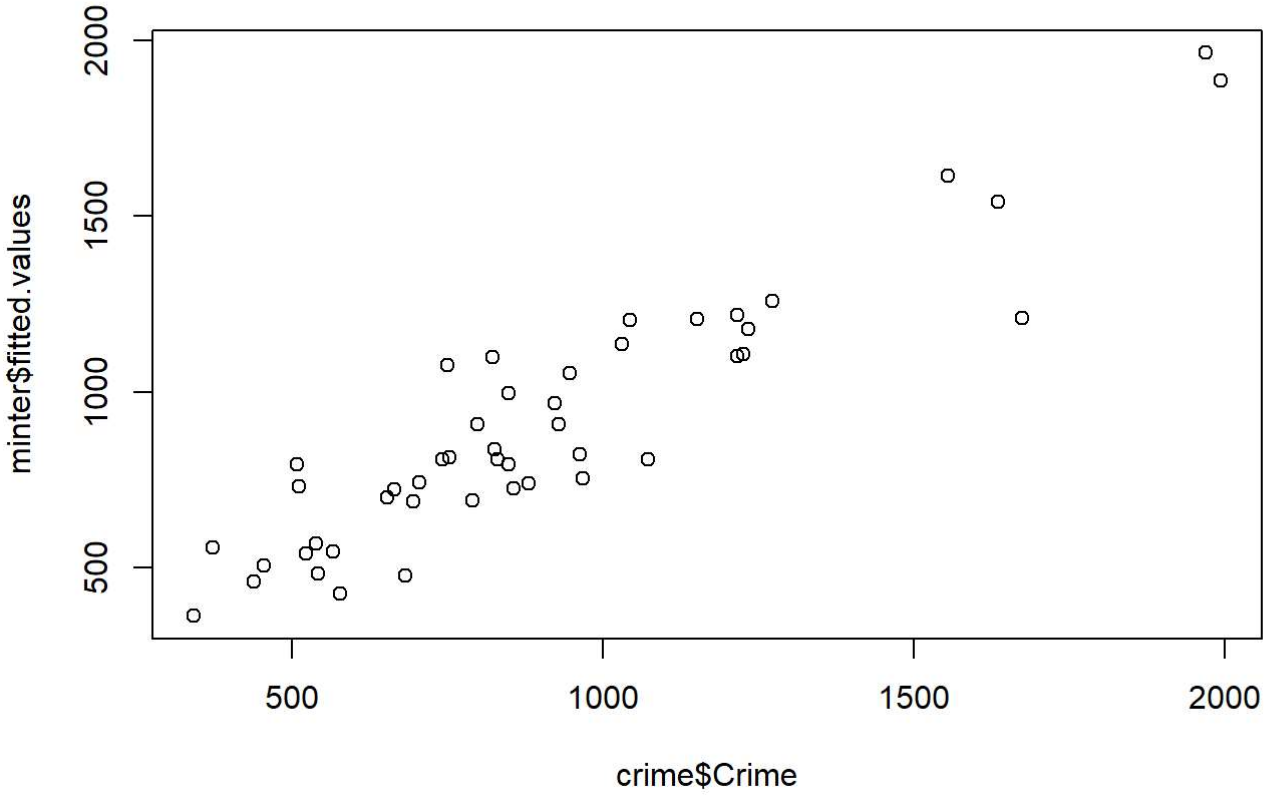
Normality is reasonable based on inspection, although points extremes may not follow a completely normal distribution. Based on Shapiro-Wilk test we can assume normal distribution of residuals.

```
plot(cooks.distance(minter), type="h")
```



In Cook's distance, the most influential point's score has reduced to about .15. Let's take a look at fitted vs. actual values with this model:

```
# Plot fitted vs. actual values
plot(x=crime$Crime, y=minter$fitted.values)
```



There is a clearer linear relationship between predicted and actual crime values, and we can see that generally this model did a good job of predicting crime rate across the values of crime. After checking assumptions, we can use the final model (With and without interaction terms) to predict crime rate for the test data point.

Step 3: Use final model to predict crime rate for test data point

```
# Read test point as dataframe:
test_scenario <- 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 crime rate using model without interaction terms:
pred1 <- predict(mbest, test_scenario)
pred1
```

```
##        1
## 1038.413
```

```
# Predict crime rate - model with interaction terms
pred2 <- predict(minter, test_scenario)
pred2
```

```
##        1
## 1217.654
```

Predicted value for crimes is 1038 for the model without interaction terms, and 1217 for the model with interaction terms. While overfitting could be a problem for any model with this small of a data set (so small it's hard to even to CV!), I think 1217 will be a better prediction as the model with interaction terms performed better in terms of R-squared, AIC, and MSE. This seems like a reasonable value, looking at the predicted vs. actual crime rates graph above. We can see that this value is well within the range of most of the other data points, and points with similar crime levels were predicted fairly accurately.

Summary:

Best model: No interaction terms

Factors Used: M, Ed, Po1, M.F, U1, U2, Ineq, Prob (8 factors).

Quality of fit:

This model had an adjusted R-squared = .744, AIC = 639, BIC = 658, and MSE = 30916. This was an improvement over the original model with all 15 predictors, which had adjusted R-squared = .707, AIC = 650, BIC = 681, and MSE = 28828 (3 out of 4 quality measures improved). This model predicted a crime rate of 1038 for the test data point.

Coefficients

```
mbest$coefficients
```

```
## (Intercept)           M          Ed         Po1         M.F          U1
## -6426.10102    93.32155   180.12011   102.65316    22.33975 -6086.63315
##          U2        Ineq        Prob
##   187.34512    61.33494 -3796.03183
```

Output:

```
summary(mbest)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.70 -111.07    3.03  122.15  483.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6426.10    1194.61  -5.379 4.04e-06 ***
## M              93.32      33.50   2.786  0.00828 **
## Ed            180.12      52.75   3.414  0.00153 **
## Po1           102.65      15.52   6.613 8.26e-08 ***
## M.F            22.34      13.60   1.642  0.10874
## U1          -6086.63    3339.27  -1.823  0.07622 .
## U2            187.35      72.48   2.585  0.01371 *
## Ineq           61.33      13.96   4.394 8.63e-05 ***
## Prob        -3796.03    1490.65  -2.547  0.01505 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195.5 on 38 degrees of freedom
## Multiple R-squared:  0.7888, Adjusted R-squared:  0.7444
## F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

Best model with interaction terms:

Quality of fit:

The best model I was able to find used the factors above, plus added the following interaction terms: U1*U2, Ed*Ineq, M*Ineq, Po1*Ineq, and Ed*Po1
Final model:
Crime ~ M+Ed+Po1+M.F+U1+U2+Ineq+Prob+U1*U2+Ed*Ineq+M*Ineq+Po1*Ineq+Ed*Po1 This model had an adjusted R-squared = .801, AIC = 631, BIC = 658, and MSE = 20816. This model predicted a crime rate of 1217 for the test data point.

Coefficients

```
minter$coefficients
```

```
##  (Intercept)            M           Ed          Po1          M.F           U1
## -4418.820561   332.231708  -172.431071  -322.461838    15.839885    58.276616
##           U2         Ineq         Prob        U1:U2      Ed:Ineq       M:Ineq
##   385.699602    82.345386 -2984.525824 -1896.690784     5.593426   -11.902060
##     Po1:Ineq       Ed:Po1
##    10.107715    23.043435
```

Output:

```
summary(minter)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob +
##     U1:U2 + Ed:Ineq + M:Ineq + Po1:Ineq + Ed:Po1, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -324.73  -59.94   -9.71   97.49  464.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4418.821   4666.594  -0.947   0.3506
## M             332.232    150.395   2.209   0.0342 *
## Ed           -172.431    295.922  -0.583   0.5641
## Po1          -322.462    231.305  -1.394   0.1726
## M.F            15.840     12.587   1.258   0.2171
## U1             58.277   6737.936   0.009   0.9932
## U2            385.700    185.881   2.075   0.0459 *
## Ineq           82.345    172.983   0.476   0.6372
## Prob        -2984.526   1345.403  -2.218   0.0335 *
## U1:U2       -1896.691   1677.696  -1.131   0.2664
## Ed:Ineq         5.593     10.959   0.510   0.6132
## M:Ineq        -11.902      7.109  -1.674   0.1035
## Po1:Ineq       10.108      4.744   2.131   0.0407 *
## Ed:Po1         23.043     16.102   1.431   0.1618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.2 on 33 degrees of freedom
## Multiple R-squared:  0.8578, Adjusted R-squared:  0.8018
## F-statistic: 15.31 on 13 and 33 DF,  p-value: 2.176e-10
```

Ideas for further exploration

I want to learn more about best practices and practical tips for variable selection! I tried doing some cross validation with this data set, but found it was so small and it was harder to get adjusted R-squared (vs. R-squared), AIC, and BIC, model coefficients, and p-values to compare models and iterate through variable selection. This is something I'd like to practice with a larger dataset.