Crime rate for 42 cities

Elena Shevchenko

Data

The data set includes information for 42 cities, and here is the description of the variables: for each city,

Y - total overall reported crime rate per 1 million residents,

X_1 - reported violent crime rate per 100,000 residents,

X_2 - annual police funding in dollars per resident,

X_3 - percent of people 25+ year-olds with 4 years of high school,

X_4 - percent of 16 to 19 year-olds not in high school and not high school graduates,

X_5 - percent of 18 to 24 year-olds in college,

X_6 - percent of people 25+ year-olds with at least 4 years of college.

Analysis

Scatterplot matrix for all variables is presented below along with correlation matrix. Y variable has high correlation with X_1 and not high but moderate correlation with X_2. Some variables have some sort of correlation between each other, meaning that we have to check multicollinearity.

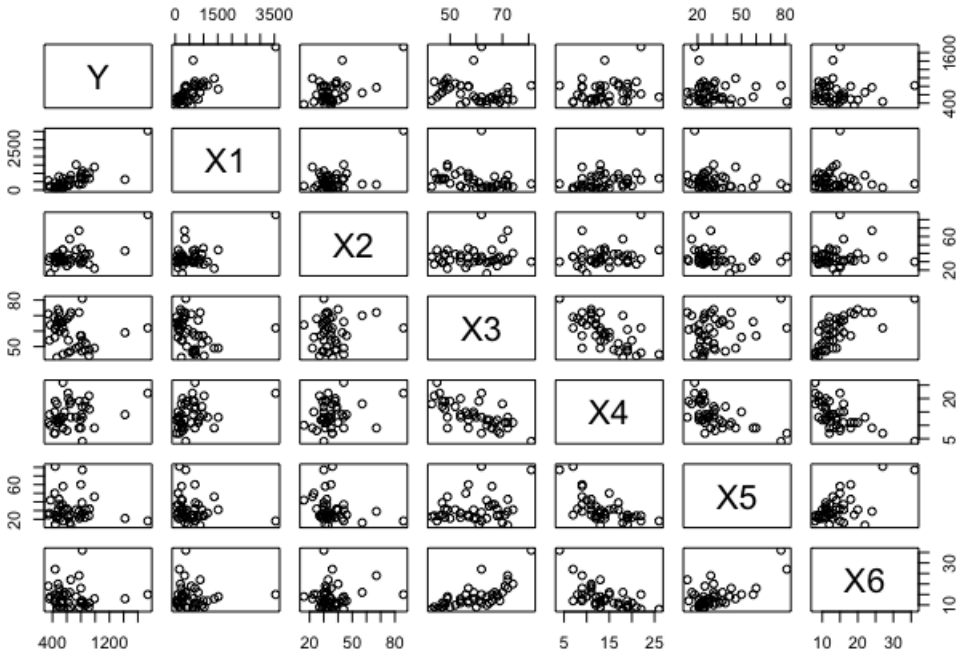

Figure 1: Scatter plot

	Y	X1	X2	X3	X4	X5	X6
Y	1.00000000	0.77683616	0.5341683	-0.2258631	0.2639391	-0.1197913	-0.02739168
X1	0.77683616	1.00000000	0.5298509	-0.2484695	0.3417110	-0.1940981	-0.05878681
X2	0.53416833	0.52985094	1.0000000	0.1158222	0.2581642	-0.2704773	0.11412139
X3	-0.22586307	-0.24846953	0.1158222	1.0000000	-0.6714258	0.2376885	0.70382437
X4	0.26393907	0.34171097	0.2581642	-0.6714258	1.0000000	-0.6132425	-0.62076104
X5	-0.11979130	-0.19409812	-0.2704773	0.2376885	-0.6132425	1.0000000	0.62007344
X6	-0.02739168	-0.05878681	0.1141214	0.7038244	-0.6207610	0.6200734	1.00000000

Figure 2: Correlation matrix

Even though Residuals vs Fitted values plot has a red curve instead of straight line, there is no clear pattern of points. It is also clear that case 38 may be influential. Normal Q-Q plot shows that all points are close to the line, except cases 12, 36 and 38, meaning that the data is approximately normally distributed. According to the Scale-location plot, the data is homoscedastic since points spread out randomly and the pattern is not clear, even though the red curve is not the straight line. The Residuals vs Leverage shows that observations 12, 36 and 38 are close to Cook's bounds. The graph for Cook's distance shown below has several points that are greater than $\frac{4}{n-p-1} = 0.11$, meaning that they can be potential influential cases. Using diagonal elements of the hat matrix, which should be less than $\frac{2p}{n} = 0.29$, it can be concluded that cases 5, 12, 18, 37, 39, 41 are potentially influential. As a result of excluding cases 12 and 38, the model becomes more significant. Hence there will be two models with all cases included and these two influential cases excluded.

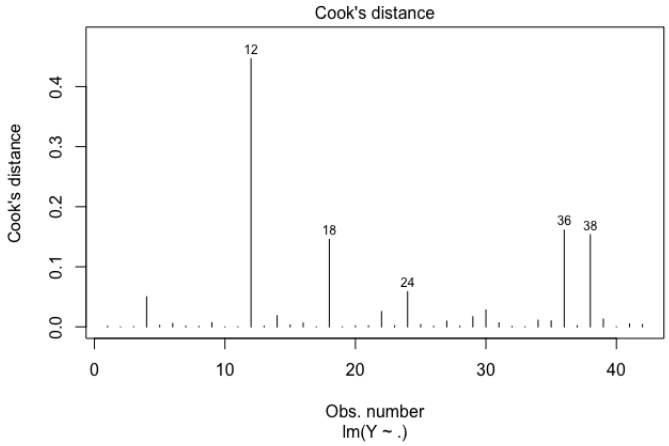

Figure 3: Cook's distance

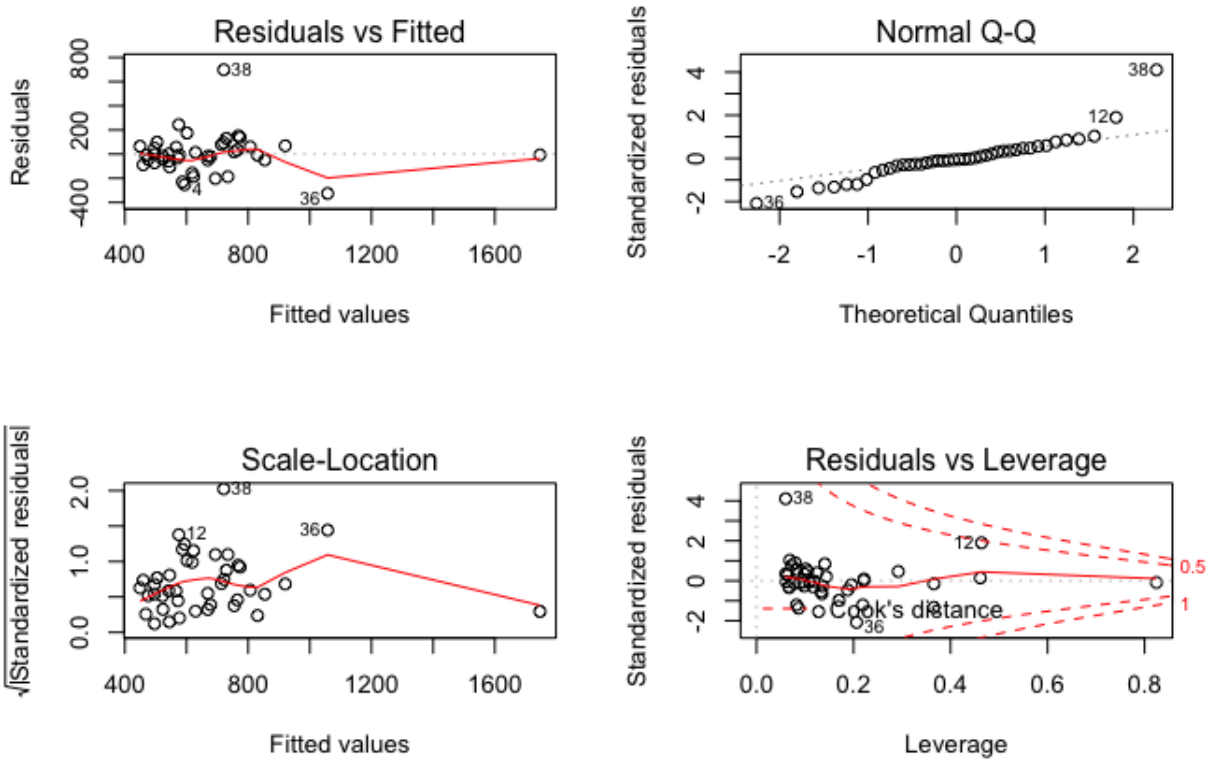

Figure 4: Diagnostic plots

Results of the linear regression without case 12 and case 38 are presented below. The model has higher adjusted coefficient of determination, equal to 0.77. Additionally, X_1 and X_2 are highly significant along with a constant. Thus, reported violent crime rate and annual police funding have significant impact on the total overall reported crime rate. The results of multicollinearity checking are presented below. There is no multicollinearity since all values of VIF are low.

```
Call:
lm(formula = Y ~ ., data = HW3_2_12_38)

Residuals:
     Min       1Q   Median       3Q      Max
-279.911  -61.385    9.093   80.912  174.541

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 625.29662  307.20509   2.035   0.0499 *
X1            0.29895    0.04056   7.370 1.83e-08 ***
X2            5.70946    2.20750   2.586   0.0143 *
X3           -3.81264    4.03276  -0.945   0.3513
X4           -5.07898    7.45453  -0.681   0.5004
X5            1.65498    2.22940   0.742   0.4631
X6           -8.32662    7.55898  -1.102   0.2786
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 119 on 33 degrees of freedom
Multiple R-squared:  0.8064,    Adjusted R-squared:  0.7712
F-statistic: 22.91 on 6 and 33 DF,  p-value: 1.858e-10
```

Figure 5: Regression summary

```
       X1       X2       X3       X4       X5       X6
1.685944 1.948349 3.832126 3.311287 2.516609 3.080642
```

Figure 6: VIF

Results of stepwise regression using backward direction show that the best model is $Y = X_1 + X_2 + X_5 + X_6$. Stepwise regression using mixed approach gives the same results. Thus, reported violent crime rate, annual police funding, percent of 18 to 24 year-olds in college and percent of people 25+ year-olds with at least 4 years of college have significant impact on the total overall reported crime rate. Increase in reported violent crime rate per 100,000 residents by 1 unit leads to increase in total reported crime rate by 0.3106. Increase in annual police funding by 1 unit leads to increase in total reported crime rate by 5.1199. Increase in percent of 18 to 24 year-olds in

college by 1 unit leads to increase in total reported crime rate by 2.9367. Increase in percent of people 25+ year-olds with at least 4 years of college by 1 percent leads to decrease in total reported crime rate by 12.0323.

```
Step:  AIC=385.75
Y ~ X1 + X2 + X5 + X6

       Df Sum of Sq     RSS    AIC
<none>               480536 385.75
- X5    1     39727  520262 386.93
- X6    1     70716  551252 389.24
- X2    1     85231  565767 390.28
- X1    1    976889 1457425 428.13

Call:
lm(formula = Y ~ X1 + X2 + X5 + X6, data = HW3_2_12_38)

Coefficients:
(Intercept)           X1           X2           X5           X6
   354.4708       0.3106       5.1199       2.9367     -12.0323
```

Figure 7: Stepwise regression

Conclusion

Reported violent crime is part of total overall crime, therefore, it is clear that it affects the response. Police funding can be related to the total overall reported crime through additional resources that police may use to be more accessible to people. Young people from college may impact overall crime rate as students contact more with other people and as a result there is higher probability that they face crime. People with at least 4 years of college may be considered as more intellectual and educated people who know the law and have higher quality of life. As a result, the more people with at least 4 years of college, the less crime is reported.