Factors affecting the murder rate in the United States

word count: 1033

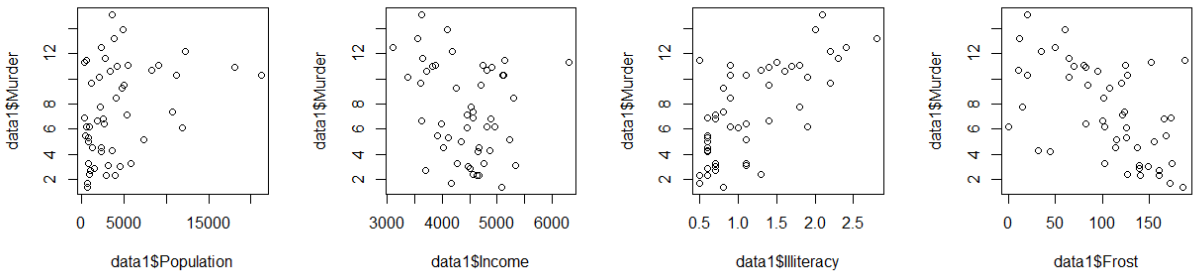
1928613

Introduction

Violent shootings in the United States are well known. However, the bill banning the sale of firearms has been delayed. Obviously, the proliferation of guns is one of the reasons for the high crime rate in all States of the United States, there are over 31000 firearm deaths in the United States in 2010, which became the highest cause of death for men under 40(Miller et al., 2013).Although Setting up laws to ban gun trafficking can greatly reduce the murder rate, but it seems very difficult to achieve at present.Therefore, predicting the city's crime rate through the characteristics of some cities, such as population, average education level and average income, helps to improve a city's control over the crime rate. Therefore, through the state dataset from Rstudio, We obtained murder rate data from 50 states in the United States, as well as other factors (population, per capital income, average illiteracy rate of the population, life expectancy, percentage of people who graduated from high school, average number of days in a year when the average temperature is less than 0 degrees, and land area in square miles)(Becker et al., 1988).This is a data table with 50 rows and 8 columns. It can be observed that all factors are continuous variables. Therefore, we can establish a regression model of regional murder rate and other factors through multiple linear regression.

Data analysis

First, We make a preliminary observation on the data set and make a scatter diagram for a single factor.



From the above figure, we can observe that the illiteracy rate of the population is slightly positively correlated with the murder rate, but the mean the average number of days when the annual average temperature is lower than 0 ℃ is slightly negatively correlated with the murder rate.Then we performed linear regression on all factors.The regression coefficient table should confirm the prediction of image observation. The former has the highest negative correlation value of -0.54 and the latter has the highest positive correlation value of 0.70

```
##            Population Income Illiteracy Murder Frost
## Population       1.00   0.21       0.11   0.34 -0.33
```

```
## Income           0.21   1.00      -0.44  -0.23  0.23
## Illiteracy       0.11  -0.44       1.00   0.70 -0.67
## Murder           0.34  -0.23       0.70   1.00 -0.54
## Frost           -0.33   0.23      -0.67  -0.54  1.00
```

After summarizing the regression model, we found that although the p value of the overall model is not significant, only the illiteracy rate is significant. Therefore, we cannot believe in a linear regression model composed of all factors. We need to analyze the data and reconstruct the model.

```
##
## Call:
## lm(formula = Murder ~ ., data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7960 -1.6495 -0.0811  1.4815  7.6210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.235e+00  3.866e+00   0.319   0.7510
## Population  2.237e-04  9.052e-05   2.471   0.0173 *
## Income      6.442e-05  6.837e-04   0.094   0.9253
## Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
## Frost       5.813e-04  1.005e-02   0.058   0.9541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 45 degrees of freedom
## Multiple R-squared:  0.567,  Adjusted R-squared:  0.5285
## F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

```
##                     2.5 %       97.5 %
## (Intercept) -6.552191e+00 9.0213182149
## Population   4.136397e-05 0.0004059867
## Income      -1.312611e-03 0.0014414600
## Illiteracy   2.381799e+00 5.9038743192
## Frost       -1.966781e-02 0.0208304170
```

-Independence test

If the error term is not independent, there will be no theoretical basis for the estimation of error term and hypothesis test. Therefore, we introduce Durbin-Watson test to judge the independence. DW is close to 0, indicating that there is a positive correlation in the residual. If DW is close to 4, it is a negative correlation. If DW is close to 2, it indicates that the residuals are independent(Tillman, 1975).

$$DW = \sum_{n}^{2}(e_i - e_{i-1})^2/ESS$$

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2006929      2.317691    0.26
##  Alternative hypothesis: rho != 0
```

We can see that the DW value is close to 2, that is, the residuals are independent of each other, and then we will carry out global regression.

- Collinearity diagnostics

Variance expansion factor (VIF) is a closely related statistic of diagnostic collinearity in multiple regression(Miles, 2014). If Vif > 5, it indicates that there is obvious correlation between variables, which will seriously affect the regression model. They are based on R square value, which is obtained by regressing the predicted value with all other predicted values in the analysis.

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}$$

```
## Population     Income Illiteracy      Frost
##   1.245282   1.345822   2.165848   2.082547
```

By solving the Vif function, the Vif values of the four factors are less than 5. Therefore, we can say that the four variables are not collinear and independent of each other, so we can further establish the regression model.

- Global Regression

Traverse the possibility of all factor combinations, and calculate the PC value, adjr2 value, and BIC value of each combination.The determination coefficient R-squared explains the extent to which the regression model can explain the differences, while adjr2 can avoid over fitting.BIC increases with the increase of model complexity, so we choose a small BIC value to reduce the complexity of the model.Comparing the four data tables, we found that the regression model formed by the two factors of crime rate and population,illiteracy rate, has the highest adjr2 value and the lowest BIC value. Therefore, we can deduce that the regression model formed by these two factors is the best.

$$R_{adj}^2 = 1 - \frac{SS_{res}/(n-k)}{SS_T/(n-1)}$$

```
## Subset selection object
## Call: regsubsets.formula(Murder ~ ., data = data1, nbest = 6)
## 4 Variables  (and intercept)
##            Forced in Forced out
## Population     FALSE      FALSE
## Income         FALSE      FALSE
## Illiteracy     FALSE      FALSE
## Frost          FALSE      FALSE
## 6 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          Population Income Illiteracy Frost
## 1  ( 1 ) " "        " "    "*"        " "
## 1  ( 2 ) " "        " "    " "        "*"
## 1  ( 3 ) "*"        " "    " "        " "
## 1  ( 4 ) " "        "*"    " "        " "
## 2  ( 1 ) "*"        " "    "*"        " "
## 2  ( 2 ) " "        " "    "*"        "*"
## 2  ( 3 ) " "        "*"    "*"        " "
## 2  ( 4 ) "*"        " "    " "        "*"
## 2  ( 5 ) " "        "*"    " "        "*"
## 2  ( 6 ) "*"        "*"    " "        " "
## 3  ( 1 ) "*"        "*"    "*"        " "
## 3  ( 2 ) "*"        " "    "*"        "*"
```

```
## 3  ( 3 ) " "        "*"    "*"        "*"
## 3  ( 4 ) "*"        "*"    " "        "*"
## 4  ( 1 ) "*"        "*"    "*"        "*"
```

```
##  [1]  6.562469 27.737968 45.642898 52.413387  1.012217  7.724108  7.797396
##  [8] 26.571510 28.457238 37.759990  3.003343  3.008879  9.106216 25.450282
## [15]  5.000000
```

```
##  [1] 0.48363609 0.27561193 0.09971722 0.03320520 0.54840003 0.48106075
##  [7] 0.48032547 0.29196796 0.27304876 0.17971582 0.53867360 0.53861685
## [13] 0.47611343 0.30857146 0.52845693
```

```
## NULL
```

AIC value is a monotonic function of the sum of squares of residuals. With the increase of the number of independent variables, the sum of squares of residuals decreases and the goodness of fit of the model becomes better(Kuha, 2004). Therefore, the goodness of fit of the model is further verified by manually calculating the values of AIC and BIC.

$$AIC = n \times \ln(\frac{SSE}{n}) + 2 \times (p+1)$$

$$BIC = n \times \ln(\frac{SSE}{n}) + \ln(n) \times (p+1)$$

```
## [1] 93.76267
```

```
## [1] 96.58671
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7652 -1.6561 -0.0898  1.4570  7.6758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
## Population  2.242e-04  7.984e-05   2.808  0.00724 **
## Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.481 on 47 degrees of freedom
## Multiple R-squared:  0.5668, Adjusted R-squared:  0.5484
## F-statistic: 30.75 on 2 and 47 DF,  p-value: 2.893e-09
```

The model with the minimum AIC value and BIC value is completely consistent with the previous model, so we can determine that it is the optimal multiple regression model.

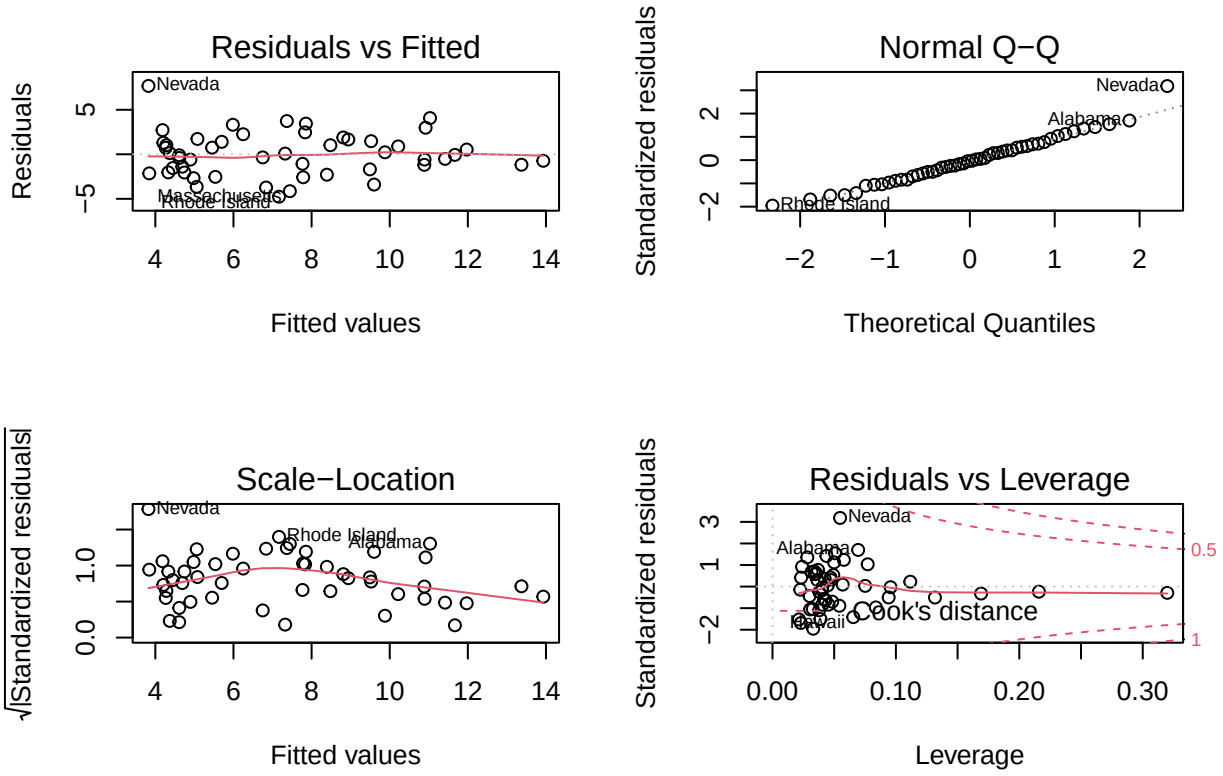
- Stepwise Regression

```
## Start:  AIC=97.75
## Murder ~ Population + Income + Illiteracy + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Frost       1     0.021 289.19  95.753
## - Income      1     0.057 289.22  95.759
## <none>                    289.17  97.749
## - Population  1    39.238 328.41 102.111
## - Illiteracy  1   144.264 433.43 115.986
##
## Step:  AIC=95.75
## Murder ~ Population + Income + Illiteracy
##
##              Df Sum of Sq    RSS     AIC
## - Income      1     0.057 289.25  93.763
## <none>                    289.19  95.753
## - Population  1    43.658 332.85 100.783
## - Illiteracy  1   236.196 525.38 123.605
##
## Step:  AIC=93.76
## Murder ~ Population + Illiteracy
##
##              Df Sum of Sq    RSS     AIC
## <none>                    289.25  93.763
## - Population  1    48.517 337.76  99.516
## - Illiteracy  1   299.646 588.89 127.311

##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7652 -1.6561 -0.0898  1.4570  7.6758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
## Population  2.242e-04  7.984e-05   2.808  0.00724 **
## Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.481 on 47 degrees of freedom
## Multiple R-squared:  0.5668, Adjusted R-squared:  0.5484
## F-statistic: 30.75 on 2 and 47 DF,  p-value: 2.893e-09

## Analysis of Variance Table
##
## Model 1: Murder ~ Population + Illiteracy
## Model 2: Murder ~ Population + Income + Illiteracy + Frost
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

```
## 1      47 289.25
## 2      45 289.17  2  0.078505 0.0061 0.9939
```

The global regression method is difficult to fully expand the regression model with multiple factors, so you can borrow the step package to model step by step.By comparing the AIC values after adding and deleting factors, we can judge the importance of deleted factors to the model. After regression, we can get a regression model which is the same as the global regression, with a R-squared value of 0.5668 and an overall p value of 2.893e-09, which is more significant.

- Regression diagnosis



From Q-Q plot, it shows that most state follow the standard distribution, but the Nevada state is quite away from the theoretical value distribution line, if we delete the outlier, the regressio model will become here.

```
## Start:  AIC=84.69
## Murder ~ Population + Income + Illiteracy + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Income      1     0.908 225.89  82.884
## - Frost       1     1.063 226.05  82.917
## <none>                    224.98  84.686
## - Population  1    47.914 272.90  92.147
## - Illiteracy  1   135.602 360.59 105.800
##
## Step:  AIC=82.88
```

```
## Murder ~ Population + Illiteracy + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Frost       1     1.020 226.91  81.104
## <none>                    225.89  82.884
## - Population  1    48.013 273.90  90.327
## - Illiteracy  1   166.587 392.48 107.953
##
## Step:  AIC=81.1
## Murder ~ Population + Illiteracy
##
##              Df Sum of Sq    RSS     AIC
## <none>                    226.91  81.104
## - Population  1     59.93 286.85  90.589
## - Illiteracy  1    334.42 561.33 123.486

##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5517 -1.6481 -0.0394  1.6659  3.9888
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.052e+00  7.446e-01   1.413  0.16447
## Population  2.505e-04  7.187e-05   3.486  0.00109 **
## Illiteracy  4.359e+00  5.294e-01   8.234 1.34e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.221 on 46 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.636
## F-statistic: 42.93 on 2 and 46 DF,  p-value: 3.03e-11
```

Through the analysis, we can know that although the overall R-squared value has increased, the p value of each factor has changed, and the intercept p value has increased, which becomes insignificant, but on the whole, the model is more optimized after deleting outlier points.From the p value, it can be found that the illiteracy rate has the most significant impact on the murder rate in a region, far exceeding the impact of population on the murder rate in a region.

Conclusion

Finally, two significant factors were selected, the population of the region and the illiteracy rate of the region (the proportion of the population graduating from middle and high schools in the region). The regression model is:

```
##  (Intercept)   Population   Illiteracy
## 1.0519769129 0.0002505023 4.3588625086
```

$$Y = 1.052 + 0.00025X_1 + 4.359X_2$$

Therefore, improving people's education level will help to greatly reduce the murder rate in cities, and an appropriate reduction in the population of a city will also help to reduce the murder rate.

Reference

Becker, R., Chambers, J., Wilks, A., 1988. The new s language. Wadsworth & brooks/cole. Computer Science Series, Pacific Grove, CA.

Kuha, J., 2004. AIC and BIC: Comparisons of assumptions and performance. Sociological methods & research 33, 188–229.

Miles, J., 2014. Tolerance and variance inflation factor. Wiley StatsRef: Statistics Reference Online.

Miller, M., Azrael, D., Hemenway, D., 2013. Firearms and violent death in the united states. Reducing gun violence in America: Informing policy with evidence and analysis 3–20.

Tillman, J.A., 1975. The power of the durbin-watson test. Econometrica: Journal of the Econometric Society 959–974.