GR5291 Final Project Report

Qiuyu Ruan (qr2127), Nuanjun Zhao (nz2295), Yufan Zhang (yz3385)

Summary

Social security is a hot topic that everyone cares about. Furthermore, social stability and crime rate are closely related. So we interest in the crime data and do research on it. Therefore, we use statistical methods to conduct our research. This report is a research on the influential factors of crime rates in the United States.

In order to study crime rate, we regress crime rate on land area per person (v4), percent of population aged 18-34 (v6), percent of the population 65 or older (v7), number of active physicians per person (v8), number of hospital beds per person (v9),percent high school graduates (v11), percent bachelor's degree (v12), percent below poverty level (v13), percent unemployment (v14), per capita income (v15), geographic region (v17) replaced by dummy variables D1, D2, and D3. After conducting stepwise regression and adding quadratic term, we finalize our model and conclude that land area per person, percent of population aged 18-34, number of hospital beds per person, percentage of people below poverty level, per capita income and geographic region are the factors that will affect the crime rate.

Finally, we conclude that strengthened education for people aged 18-34, an improvement for medical service, a decent job and better government subsidy for people in poverty can reduce crime rate and improve public safety. Besides, a good public safety administration in Southwest region is important to lower crime rate.

Introduction

Data Description

The data we use is a crime-related dataset with the sample size of 440. It contains identification number(v1), county name(v2), state(v3), land area(v4), total population(v5), percent of population aged 18-34(v6), percent of the population 65 or older(v7), number of active physicians(v8), number of hospital beds(v9), total serious crimes(v10), percent high school graduates(v11), percent bachelor's degree(v12), percent below poverty level(v13), percent unemployment(v14), per capita income(v15), total personal income(v16) and geographic region(v17).

Research Objective

Crime rate is one of the most important factors in evaluating whether a county is livable. So it is meaningful to do the research regarding crime data. Through studying crime

data, we can find out the key factors that influencing crime rate in a specific area. Therefore, targeted suggestions can be made according to the key factors that are relevant to crime rate.

Initial Data Exploration

We randomly select 300 data from the raw data and do data processing. Firstly, we use the number of serious crimes divided by total population to get the crime rates, which is our object of study. In order to analysis variables in same scale, we divide some variables, v4, v8, v9, and v16, by total population to eliminate the effect and then get variables for per person and replace the old by those new variables respect to per person. We also get rid of identification number, county name and state, since they are not considerable variables and state can be replaced by region. We delete the total personal income per person v16 because it represents the same thing with per capita income v15. The geographic region v17 represents four different regions, so we make a transformation: defining dummy variables D1, D2 and D3, where D1=1, D2=0, D3=0 when v17=1 (region NE), D2=1, D1=0, D3=0 when v17=2 (region NC), D3=1,D1=0,D2=0 when v17=3 (region S). If all Ds are equal to zero, the region is W.

Results

Analysis

First, we compute the variance and covariance matrix of the independent variables, from v4 to v15. And then, we compute the VIF (Variance Inflation Factor) of these 12 variables to obtain whether the collinearity exists. Generally, when VIF is larger than 10, serious multicollinearity exists. Based on our calculation, all the VIFs are lower than 10, thus we conclude that there is no serious multicollinearity.

Second, we conduct the multiple linear regression for v4, v6 to v9, v11 to v15, and D1 to D3 respect to crime rates (y). By using stepwise regression, only 6 factors are kept according to t-test, which tests the significance of the coefficients. These 6 factors are v4, v8, v13, v15, D1 and D2. However, the R-Square and adjusted R-Square of this regression is 0.4155 and 0.4035 respectively, which is pretty small. Then, we continue our study. We check the four basic assumptions of linear regression model, which are constant variance of error terms, normality of error terms, independence of error terms and linearity.

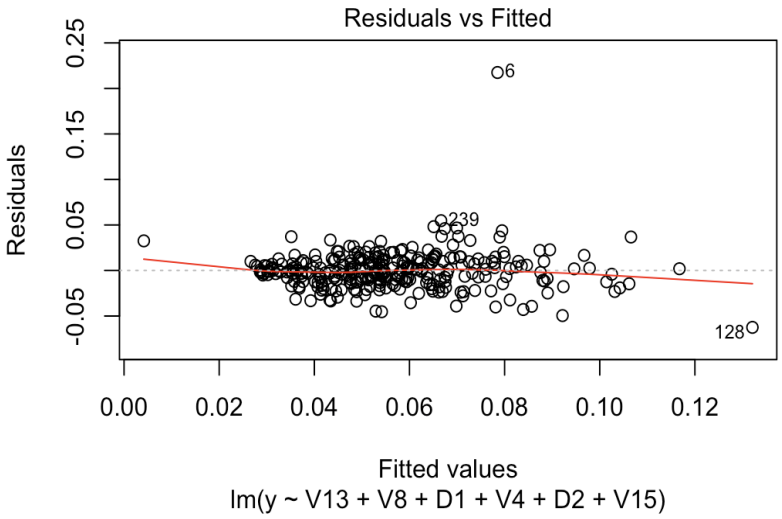

Figure 1

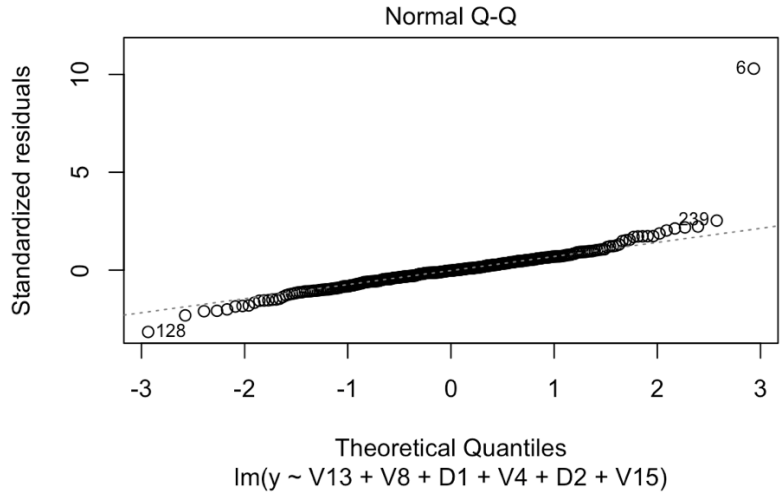

Figure 2

According to Figure 1, residuals against fitted value of y and Figure 2, normal QQ plot of residuals, there is one outlier, which is King county in NY. Also Figure 2 shows that residuals do not seem like normal distribution. So we delete the outlier and conduct multiple linear regression again using the new dataset without the outlier.

Again, using the same method as above, we conduct the multiple linear regression for v4, v6 to v9, v11 to v15, and D1 to D3 respect to crime rates (y).

The retaining variables are as follows:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.982e-03  1.159e-02  -0.861 0.389758
V13          2.030e-03  3.352e-04   6.057 4.28e-09 ***
D1          -2.720e-02  2.697e-03 -10.088  < 2e-16 ***
D2          -1.720e-02  2.544e-03  -6.759 7.57e-11 ***
V4          -5.446e-01  1.547e-01  -3.520 0.000501 ***
V9           3.341e+00  6.447e-01   5.183 4.10e-07 ***
V15          1.580e-06  3.487e-07   4.532 8.56e-06 ***
V6           7.379e-04  2.417e-04   3.053 0.002476 **
```

To check the four basic assumptions regarding linear regression, we plot four residual plots.

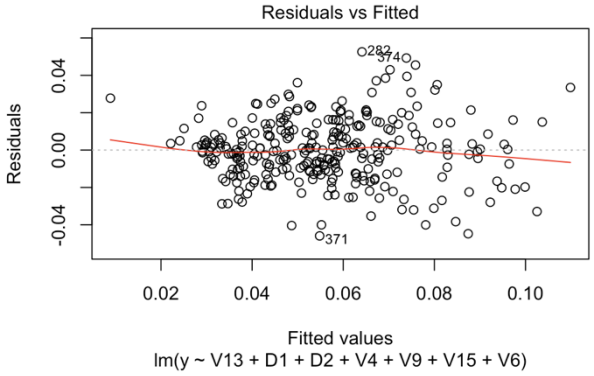

Figure 3

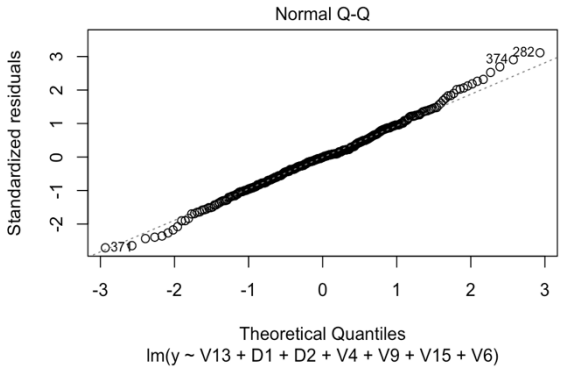

Figure 4

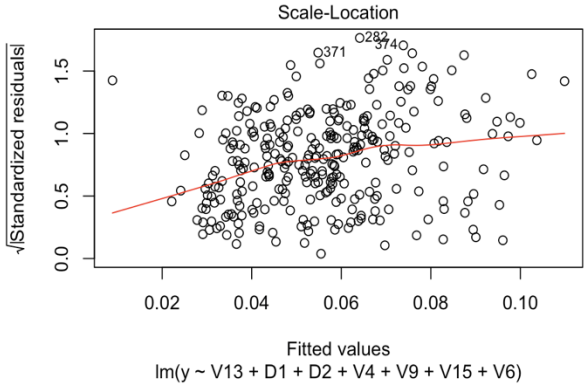

Figure 5

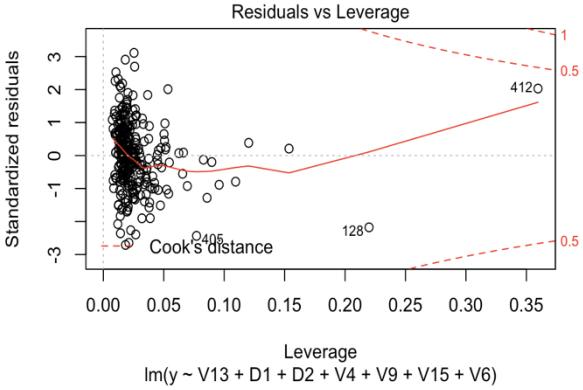

Figure 6

We decide to add quadratic terms to improve the accuracy of the model. The model using stepwise regression contains v4, v6, v9, v15, D1, D2, $v4^2$, $v9^2$, $v15^2$. The value of R-Square and adjusted R-Square are 0.5916 and 0.5774 respectively.

We finally conclude that crime rates are affected by these seven factors, v4, v6, v9, v13, v15, D1 and D2. The coefficients for the terms are as follows:

Variables	Intercept	V4	V6	V9	V13	V15
Coefficients	-0.134	-1.173	0.000969	7.718	0.002983	0.00001188
Variables	V4^2	V9^3	V15^2	D1	D2	
Coefficients	16.47	-666.90	-2.36E-10	-0.02726	-0.01734	

Table 1

Interpretation

Our model's intercept is -0.134, which represents the expected mean value of y (crime rate) when all variables are equal to zero.

For percent of population aged 18-34 (v6), its coefficient is 0.000969, which represents that a unit increase in v6 leads to 0.000969 increase in crime rate (y) while keeping other variables fixed.

For percent below poverty level (v13), its coefficient is 0.002983, which represents that a unit increase in v13 leads to 0.002983 increase in crime rate (y) while keeping other variables fixed.

For land area (v4), it has both the linear term and quadratic term, of which the coefficients are -1.173 and 16.47 respectively. -1.173 means that the rate of change of y when v13 is equal to zero while keeping all the other variables fixed. The coefficient of quadratic term represents the direction and steepness of the curvature. 16.47 is positive, which means that the curvature is upwards.

For number of hospital beds (v9), it has both the linear term and quadratic term, of which the coefficients are 7.718 and -666.9 respectively. 7.718 means that the rate of change of y when v9 is equal to zero while keeping all the other variables fixed. The coefficient of quadratic term represents the direction and steepness of the curvature. -666.9 is negative, which means that the curvature is downwards.

For per capita income(v15), it has both the linear term and quadratic term, of which the coefficients are 0.00001188 and -2.36e-10 respectively. 0.00001188 means that the rate of change of y when v15 is equal to zero while keeping all the other variables fixed. The coefficient of quadratic term represents the direction and steepness of the curvature. -2.36e-10 is negative, which means that the curvature is downwards.

For D1, which represents north east region, the coefficient is -0.02726. It tells us that the crime rate in north east region is lower than average. The lower crime rate results from

bad weather conditions. Especially in winter, the outside activities are rare due to the freezing weather.

For D2, which represents NC region, the coefficient is -0.01734. It tells us that the crime rate in NC region is lower than average. The reason why north central region has overall lower crime rate might be the lifestyle. Since agricultural is the dominated industry in north central part of the United States, people living there have easy lifestyle and less pressure, which will reduce their criminal behavior.

D3 represents south region. While D3 does not exist in our regression function, it shows that crime rates in south and west region do not have significant difference and these two regions can be regarded as a baseline. So, the crime happens in west and south does not depend on region.

Conclusion

After analysis, we know the main factors that will influence crime rate. In order to lower the crime rate, we should concentrate on land area per person (v4), percent of population aged 18-34 (v6), number of hospital beds (v9), percentage of people below poverty level (v13), per capita income (v15) and geographic region (v17).

A good education and propaganda for consciousness of legality for young people are essential to reduce the incidence rate of crime among people aged 18-34. Also, income is a sensitive factor for crime rate. A good job and better governmental subsidy can help people below poverty level improve their living quality, eventually lower the crime rate. A good medical in US can also help to reduce crime rate. The last influential factor is Southwest region. To reduce the crime rate in that region, government should focus on public security administration.