GR5291 Advanced Data Analysis (Spring 2018)
Final Report

Mingyu Jiao - mj2823
Fan Yang - fy2232
Qianhui Sun - qs2179

May 3, 2018

1 Summary

The goal of this project is to study which factors influence the number of serious crimes per county. The analysis shows that *Percent of population aged 18-34*, *Percent below poverty level*, *Per capita income*, *Percent of hospital beds*, *Population density* and *Region* are significant factors. Although *Region* is important in the analysis, there is nothing can be done to improve this factor. While the other factors all have a positive relationship with *crime rate*. Specifically, the results show that the higher level of these factors, the more likely for this county to have serious crime rate.

Based on the analysis, our advice for reducing the number serious crimes in their county is from three aspects: medical, population and salary. In other words, increasing the minimal salary, improving medical support and controling population density.

2 Introduction

2.1 Description of the data

The data we use is the county demographic information (CDI) data set from *Applied Linear Statistical Models, 5th edition, by Kutner, Nachtsheim, Neter, and Li., Appendix C2.* This data set provides selected county demographic information for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set.

As the population differ from county to county, accumulation number is not as persuasive as percentage. Therefore rescaling some variables is needed. For example, we divide *Total*

population by *Land area*; divide *Total serious crimes*, *Number of active physicians*, *Number of hospital beds* and *Number of hospital beds* by *Total population*. After that, in order to have the same scale, we reduce these variables by 10 times. Finally, all those data roughly fall into the range $[10, 500]$.

2.2 Value/Benefit of the project

The purpose of this project is to find the cause of serious crime rate. So based on our findings, government can determine which factors influence the number of serious crimes and implement policies that will lead to the reduction of the number serious crimes in their county. Specifically, crime rate in Kings county (Brookly) is in top level among all counties. Officials is thereby worried about their crime rate and help to reduce crime rate in Kings county is also purpose of this project.

Based on our project goal, we try to dig out the most significant factors among all variables in the dataset. After rescaling, we set pool of factors include 12 variables and one of them is categorical variables which is *Geographic region*. And rest of them are numerical variables. Then we try to find out the significant influential factors.

2.3 Initial data exploration

The subset that we use in the analysis consists of 12 independent variables. *Geographic region* is the only categorical variables which has four categories (1=NE, 2=NC, 3=S, 4=W). 23.41% of the counties are NE; 24.55% are NC; 34.54% are S and 17.5% are W. We first explore whether region matters for crime rate(crime rate here is percentage instead of accumulation number).

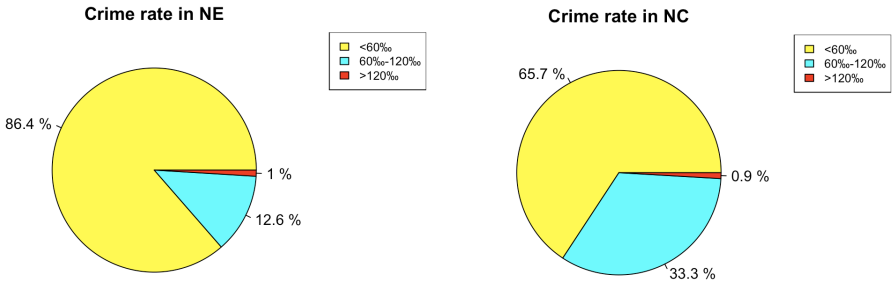

Figure 1: Crime rate in region NE and NC

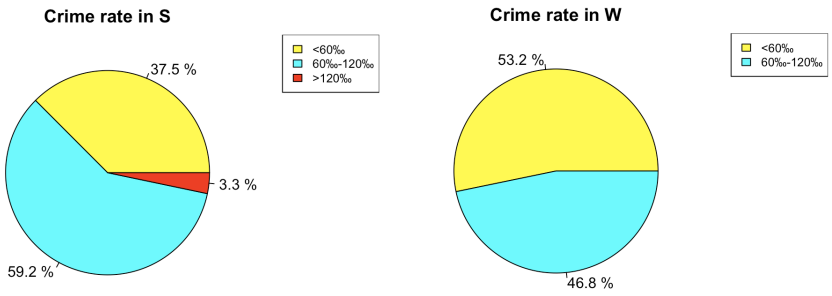

Figure 2: Crime rate in region S and W

Of those counties in S, 20.4% of them have crime rate above 90‰, while those counties in NE NC and W, less than 5% have crime rate above 90‰. In the view of plot, crime rate in different regions differ greatly and counties in S tend to have severe crime rate.

Besides *Geographic region*, we also find that *Population density* and *Percent below poverty level* important to explain high crime rate. Here we highlight the Kings county to compare its *Population density* and *Percent below poverty level* with all other counties. This is reasonable since crime rate in Kings county is much higher among all.

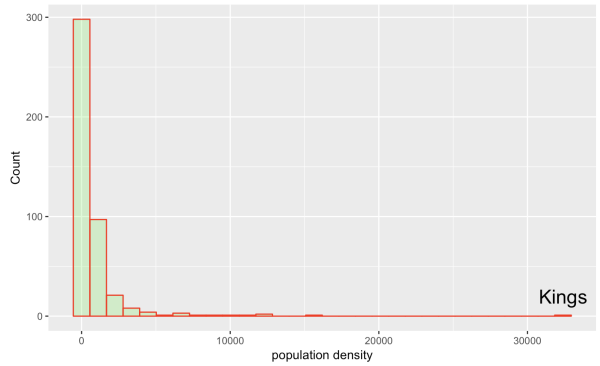

Figure 3: Population density histogram

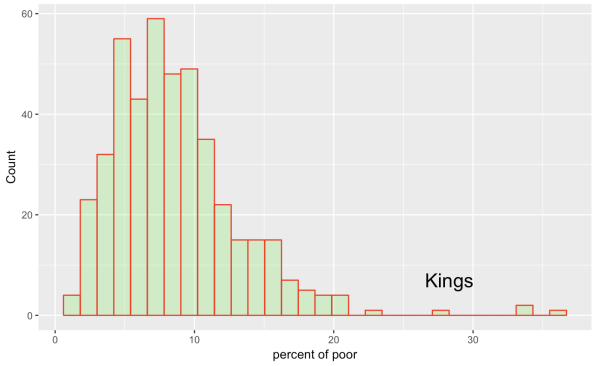

Figure 4: Percent of poor histogram

The average population density is around 888 per square miles, while in Kings county, the population density is above 30000 per square miles, much higher than most of the counties. The average percent below poverty level is around 9%, while in Kings county, it is about 28%, which is 2 times more than the average level.

3 Analysis and Results

We use three models to to analyze this data: regular linear regression, poison regression and logistic regression. The results are slightly different in the three models but the significant factors which stand out in the three models are the same. The analysis shows that *Percent of population aged 18-34*, *Percent below poverty level*, *Per capita income*, *Percent of hospital beds*, *Population density* and *Region* are significant factors.

3.1 Linear Regression

To analyze the influence factors of crime rate, the linear regression is the first choice. If the p-value of an independent variable is less than confidence level(0.05), we can say this variable has significant effects on crime rate generally. We assume that those variables follow normal distribution and independent variables have linear relation with crime rate.

At first, we simply use the full linear regression model to analyze the data. Before fit model, we use Variance inflation factor(VIF) method to check the collinearity. The higher value of VIF, the more likely this variable is related with others. Therefore we drop the *Percent bachelors degree* since it has high VIF(greater than 5).

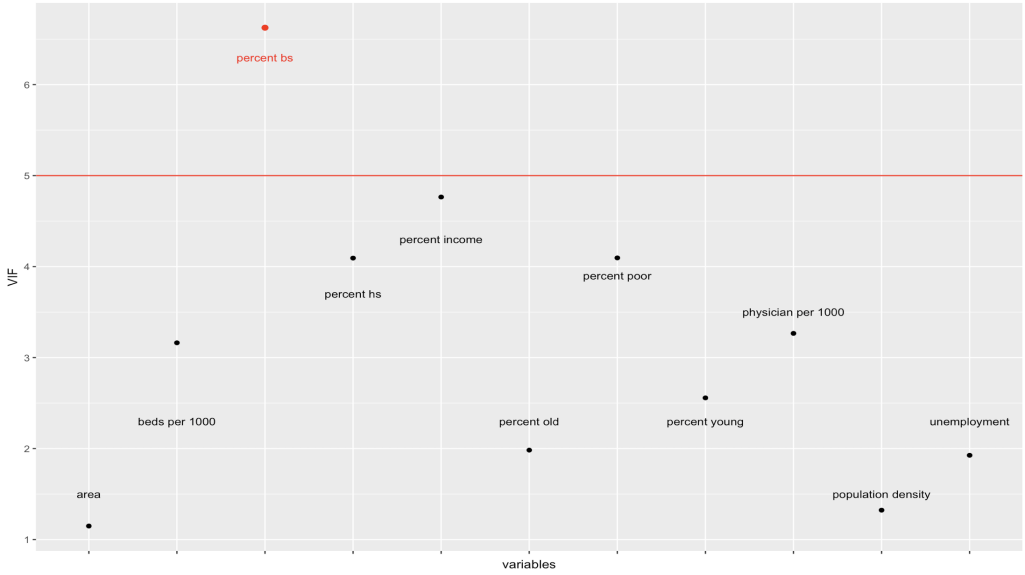

Figure 5: VIF plot

Then we run the regression with the rest ten independent variables. In order to obtain a better model, we need to do variable selection. Use backward AIC to achieve this. The AIC for full model is 2577.49 and AIC for the best reduced model is 2570.99. And least AIC can be reached when drop *percent of old*, *percent of income*, *unemployment*, *area*. The regression result shows that under 0.05 confidence level, there are only 5 variables are significant. They are *percent of young*, *perc of poor*, *percent of hospital beds*, *geographic region*, *population density*. So we can get the final model and the estimated coefficients are shown in Table 1.

		percent young	percent poor	percent beds	region	population density
estimated coefficients		0.689	1.674	2.60	9	0.0046
confidence interval	lower 2.5%	0.2713	1.1080	1.5946	7.1987	0.0037
	upper 97.5%	1.1068	2.2398	3.5957	10.8158	0.0055

Table 1: Estimated coefficient table

From the analysis, we can get the interpretation of the parameter of the variable. Take perc.young as an example, if we increase the percent of population aged 18-34 by one, the crime rate per 1000 people will increase by about 0.689. And also we can get the confidence interval for those variables. Still take perc.young as an example, we are 95% confident that a one unit increase in the percent of population aged 18-34 will increase on average the crime rate per 1000 people by a number between 0.271 and 1.107.

3.2 Poison Regression

Poisson regression is used to model response variables that are counts. It tells you which explanatory variables have a statistically significant effect on the response variable. Its best used for rare events, as these tend to follow a Poisson distribution. Poisson regression assumes the response variable has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. Poisson regression may also be appropriate for rate data, where the rate is a count of events divided by some measure of that unit's exposure. Thanks to the assumption, Poisson regression which could describe many real-world problems perfectly. So, we try this model to analyze this crime problem.

Before fit a Poisson regression, we use Pearson correlation to detect highly correlated variables.

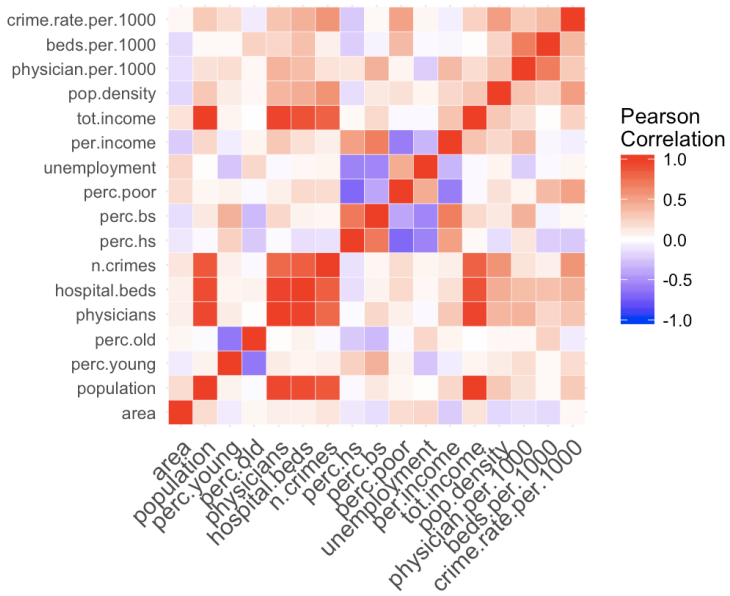

Figure 6: Pearson correlation heatmap

In Figure 6, the darker the color, the higher they are correlated. And we firstly remove those highly correlated variables. After filtration, we have the following nine variables: *percentage of young*, *percent of poor*, *percent of income*, *region*, *population*, *percent of physician*, *percent of beds*, *percentage of bachelor degrees*, *percent of unemployment*.

In the full model, there are two parameters that are not significant: young and bachelor degrees. Therefore, we refit our model without these two variables. Meanwhile, we try to detect the influence of outliers. Here we use leave-one-out deletion diagnostics. Hatvalue is one of the most common measure of leverage.

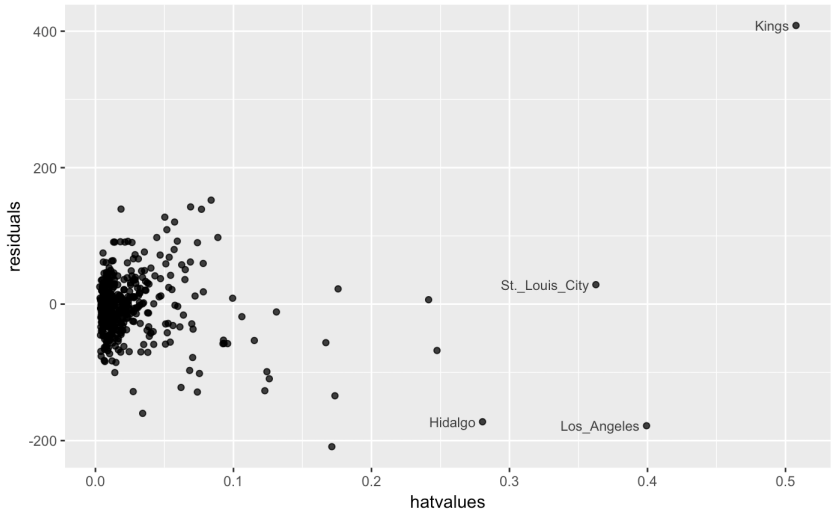

Figure 7: Residuals against Hatvalues

Based on rule of thumb, hatvalues exceeding about twice the average should be considered noteworthy. Therefore in Figure 7, we find four outlier county: *Kings*, *St. Louis city*, *Hidalgo* and *Los Angeles*. We believe that our data set has some special properties: every region only appears once. So, in this problem, we should split outliers and analyze more about them.

As discussed in initial data exploration, further from Figure 3 and Figure 4, *St. Louis city*, *Hidalgo* and *Los Angeles* shows the same pattern as *Kings*. We find that the population density and percent of poor in these four counties is much higher than others. Therefore population density and percent of poor are important reasons for outliers and the high crime rate can also be explained.

After remove outliers, we refit our model and draw the residual plot:

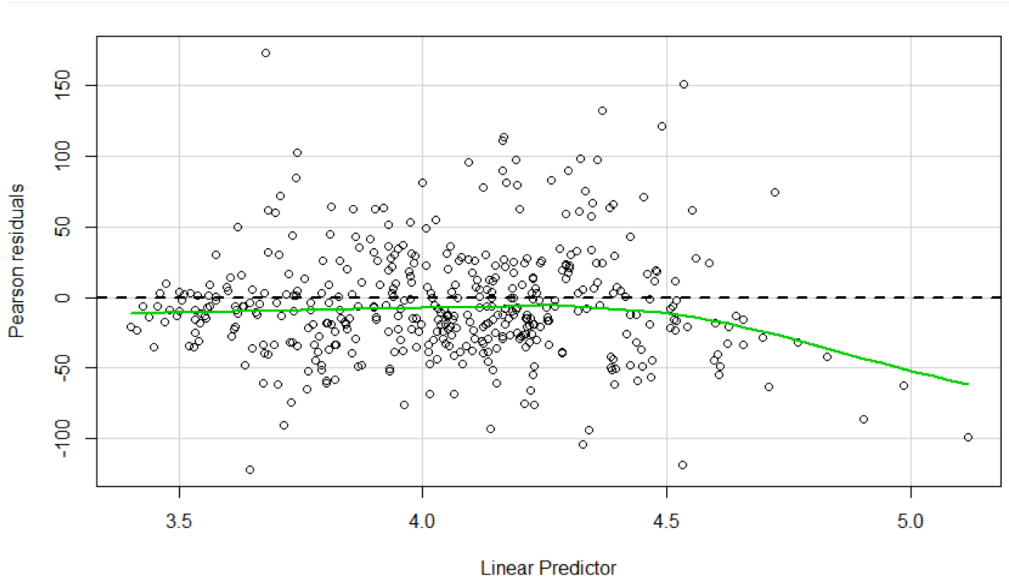

Figure 8: Residual plot

Points in the plot are randomly dispersed around the horizontal axis. Therefore our regression model is appropriate for the data. The following table gives the estimated coefficients:

	percent poor	percent income	percent beds	region	population density	percent physician
coefficients	0.0312	-0.00002	0.0286	β^1	0.2106	-0.0427
exp coefficients	1.0317	0.9999	1.0290	β^2	1.2343	0.9582

β^1: baseline category is region 1, estimated coefficients are $\beta_2^1 = 0.0506, \beta_3^1 = 0.3462, \beta_4^1 = 0.3477$
β^2: exponentiated coefficients are $\beta_2^2 = 1.0519, \beta_3^2 = 1.4137, \beta_4^2 = 1.4158$

Table 2: Estimated coefficient table

From the analysis, we can get the interpretation of the parameter of the variable. Take percent of poor as an example, if we increase the percent below poverty level by one, the crime rate per 1000 people will times 1.0318. And decreasing the percent of hospital beds by one, the crime rate per 1000 people will divide 1.029 times.

3.3 Logistic Regression

The variable concerned is crime rate, a quantitative variable. The most weakness for quantitative variable is that it is hard to tell the crime rate of a certain number is whether severe or not. Therefore, set several threshold and convert crime rate to ordinal categorical variable. Therefore, we divide the crime rate into five categories: low, medium and high. The three categories are distinguished by two cutpoints: 60‰ and 120‰ correspondingly.

Cumulative logistic regression models are used to predict an ordinal response, and have the assumption of proportional odds. Proportional odds means that the coefficients for

each predictor category must be consistent across all levels of the response. Therefore, a cumulative logit model is a good fit to our problem. The assumptions and conditions for the model are meeted. In this project, we assume different slopes between models and it will come out two models.

Before fit a cumulative logit model, we use Variance inflation factor(VIF) method to check the collinearity. The higher value of VIF, the more likely this variable is related with others. Therefore we drop the *Percent bachelors degree* since it has high VIF(greater than 5).

For the model selection process, we use backward AIC to select variables. The AIC of full model using all predictor variables 864.73. After remove *percent of old percent high school graduates* and *percent of physician*, we reach the lowest AIC 859.70. And this is our final logistic model.

In this part of analysis, we finally fit the final cumulative logit model that includes five predictor variables: *percent of young*, *percent of poor*, *percent of income*, *region* and *population density*. The estimated coefficients are shown in Table 3.

		percent young	percent poor	percent income	percent beds	region	population density
model j=1	coefficients	-0.0783	-0.1681	-0.0001	-0.1686	-0.5341	-0.00008
	exp coefficients	0.9247	0.8453	0.9999	0.8448	0.5862	0.9999
model j=2	coefficients	0.0028	-0.0660	0.00008	-0.7129	-0.3486	-0.0009
	exp coefficients	1.003	0.9362	1.000	0.4902	0.7056	0.9990

Table 3: Estimated coefficient table

Change of value of these explanatory variables implies a significant inluence on the crime rate. For example, when comparing odds ratio of $y = 1$ and $y = 2$, one unit increase of *percent of poor*(i.e.percentage increases by 1%) leads to 0.85 times of the ratio, which means a lower proportion of low crime rate to high crime rate. One unit increase of *percent of young* leads to 0.92 times of the probability ratio and one unit increase of *percent of beds* leads to 0.84 times of the ratio.

4 Conclusion

The results from the three models are almost similar except some minor difference. While *percent of income* is significant in logistic and poisson model but is removed in linear model. *Percent of physicians* only stands out in poisson model but not significant in other two. However, the three models all generally point out the following four variables: *percent of poor*, *percent of hospital beds*, *geographic region* and *population density*. Actually, *geographic region* is not changable. In the view of government, we can do nothing to it. But for citizens, based on our analysis, they can choose to live in the region that has lower crime rate, for example region in NE.

As for the rest variables, they can be grouped into three types: Medical support, Salary level and population density. Since the population density has significant effects on crime rate, government could decrease population density in some areas such as Kings, Los Angles, and control population density in other areas. For example, Governments could encourage some citizens to move to other regions with subsidy. However, in the real world, high population density means metropolis and high development rate in economy. Therefore, there is a tradeoff between economic development and crime rate.

Medical level is also a very important factors, and government should increase the investment on medical system. Government can encourage the development of health insurance, or give intervention and price control in medical costs. In our analysis, specifically, increasing the number of hospital beds is a useful method. Following this, increasing the number of hospitals and hiring more doctors and medical workers are good way to this aspect.

Last but not least, poverty plays an fatal role in influencing crime rate. Increasing social welfare and increasing the minimal salary level should be considered when implement policy. Federal unemployment insurance, federal welfare programs, and medicare, all help poor and temporarily hard-pressed households make ends meet. When people's rights and lives are guaranteed, the crime rate will decrease naturally.

5 Appendix

5.1 Best linear model

$$\begin{aligned} crime\ rate = -23.83 &+ 0.689 && \times perc.young \\ &+ 1.674 && \times perc.poor \\ &+ 2.60 && \times beds.per \\ &+ 9 && \times I(region) \\ &+ 4.59 \cdot 10^{-3} && \times pop.density \end{aligned} \tag{1}$$

5.2 Best Poisson model

$$\begin{aligned} \frac{E(Y|X)}{\text{exposure}} &= e^{\theta X} \\ \theta X = 7.993 &+ 0.0312 && \times perc.poor \\ &- 0.00002 && \times per.income \\ &+ 0.0286 && \times beds.per \\ &+ 0.0506 && \times region2 \\ &+ 0.3462 && \times region3 \\ &+ 0.3477 && \times region4 \\ &+ 0.2106 && \times pop.density \\ &- 0.0427 && \times perc.physician \end{aligned} \tag{2}$$

5.3 Best logistic model

$$
\begin{aligned}
\frac{P(Y \leq 1)}{1 - P(Y \leq 1)} &= e^{\beta X} \\
\beta X = 7.993 &- 0.0783 \quad \times perc.young \\
&- 0.1681 \quad \times perc.poor \\
&- 0.0001 \quad \times per.income \\
&- 0.1686 \quad \times beds.per \\
&- 0.5341 \quad \times I(region) \\
&- 0.00008 \quad \times pop.density \\
\frac{P(Y \leq 2)}{1 - P(Y \leq 2)} &= e^{\beta X} \\
\beta X = 7.401 &+ 0.0028 \quad \times perc.young \\
&- 0.0660 \quad \times perc.poor \\
&- 0.00008 \quad \times per.income \\
&- 0.7129 \quad \times beds.per \\
&- 0.3486 \quad \times I(region) \\
&- 0.00009 \quad \times pop.density
\end{aligned}
\tag{3}
$$

Where $\beta_1 = 7.457$, $\beta_2 = 10.72$, $\beta_3 = 13.56$, and $\beta_4 = 16.38$
Y represents the level of crime rate.