Columbia University

Graduate School of Arts and Science

Statistics Department

Advanced Data Analysis GR5291

Crime Data Analysis

STUDENT Name and ID :
Rui Liu rl2987
Anyi Wang aw3141
Yucheng Zhang yz3546

Contents

1 Summary

The goal of this project is to **explore which factors influence the number of serious crimes per county**. In the project, we used the **CrimeData**. And we determine the crime level by taking the quantile of the log of crime ratio in each county and fitting the **logistic regression**.

The final result shows that **the population density**, **the interaction of the hospital beds per area and below poverty rate** and **per capital income** have the highest influence on a county's crime level.

Therefore, the government should focus on **decreasing the population density**, **below poverty rate**, **increasing hospital beds per area and per capital income** in order to decrease crime number.

2 Data description and exploratory data analysis

2.1 Data description

The data used in this project was obtained from the real life. It contains selected demographic information for 440 of the most populous counties in the United States, which includes the estimated number of population, the total number of serious crimes, the number of professionally active non-federal physicians and other 14 variables.

Variable Name	Description
Identification Number	1-400
County Name	County name
State	Two-letter state abbreviation
...	...
Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census,where: 1=NE,2=NC,3=S,4=W

Table 1: original variable table

2.2 Exploratory Data Analysis

2.2.1 Define response and predictors

1.Response

We are interested in the association between the crime rate and the other variables.

First, we calculate the crime rate using the variable 'total serious crimes' and 'total population'.

$$crime\ rate = \frac{total\ serious\ crimes}{total\ population}$$

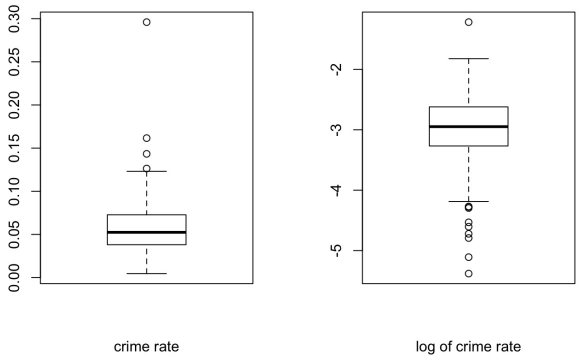


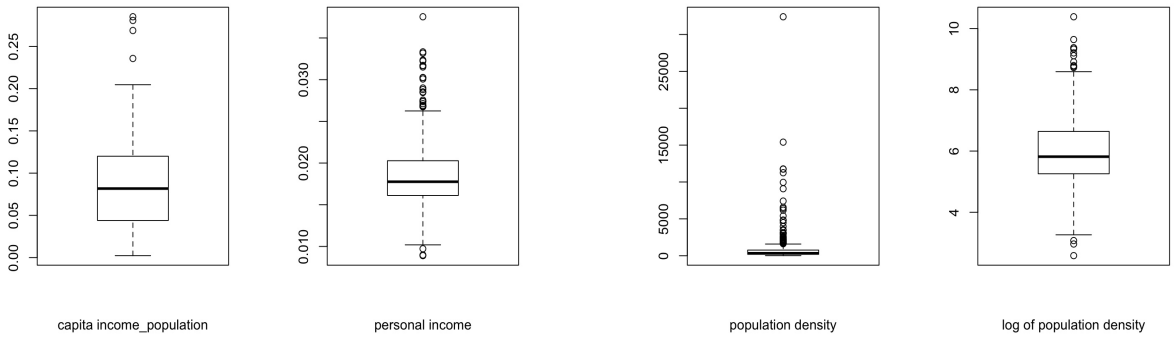
Figure 1: box plot of crime rate

Then, from the above box-plot, it shows that the crime rate is right-skewed. Therefore, we transform the variable crime rate using log function, and define the crime rate levels based on the values of 1^{st}**, median, and** 3^{rd} **quantities of the log(crime rate)**(1,2,3,4)

2.Predictors

Based on the definitions of variables, we do the following transformations:

i. income variables and density



(a) income variables (b) density

Figure 2: Plots of income and density

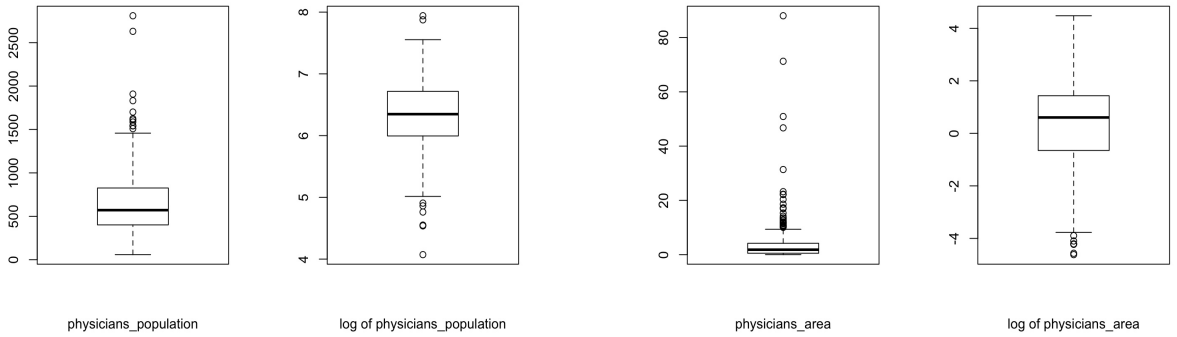
For the two income related variables, since the total income and per_capita income are roughly symmetrically distributed, we don't need to transform them.

$$population\ density = \frac{total\ population}{Landarea}$$

Based on the box-plot, we would use the log transformed values of the variable "population density" in the model.

ii. physicians

$physicians\ population = \frac{physicians}{Total\ population}$ $physicians\ area = \frac{physicians}{Landarea}$



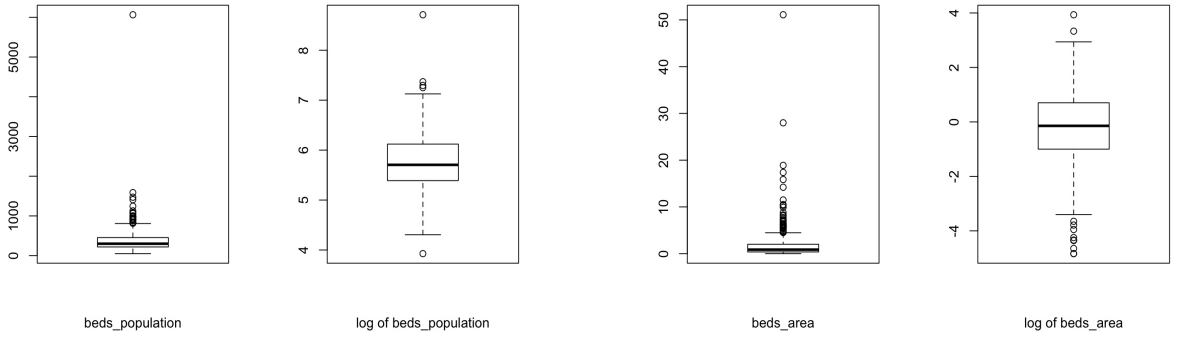
(a) physicians population (b) physicians area

Figure 3: physicians per population area

Based on the box-plot, we would use the log transformed values of them in the model.

iii. hospital beds

$hospital\ beds\ population = \frac{hospital\ beds}{Total\ population}$ $hospital\ beds\ area = \frac{hospital\ beds}{Landarea}$



(a) hospital beds population (b) hospital beds area

Figure 4: hospital beds per population area

Based on the box-plot, we would use the log transformed values of them in the model.

iv. predictors

In conclusion, our final model predictors are:

Variable Name	Description
teen_pop	population ratio of aged 18-34
elder_pop	County name
log_phy_per_area	log of the area divided by physicians number
log_bed_per_pop	log of the population divided by hospital beds
bachelor	percent of adult population (persons 25 years old or older) with bachelor's degree
high_school	percent of adult population (persons 25 years old or older) who completed 12 or more years of school
below_poverty	poverty level
log_density	log of population density
per_capita_income	Per capita income of 1990 CDI population (dollars)
income	Total personal income of 1990 CDI population (in millions of dollars)
geo_region	Geographic region classification is that used by the U.S. Bureau of the Census,where: 1=NE,2=NC,3=S,4=W

Table 2: predictor variable table

2.2.2 Correlation

The first exploratory data analysis we perform is the correlation test between variables. To proceed, we draw the Pearson Correlation Heat Map to visualize the relationship between the variables that we are going to use in the following analysis.

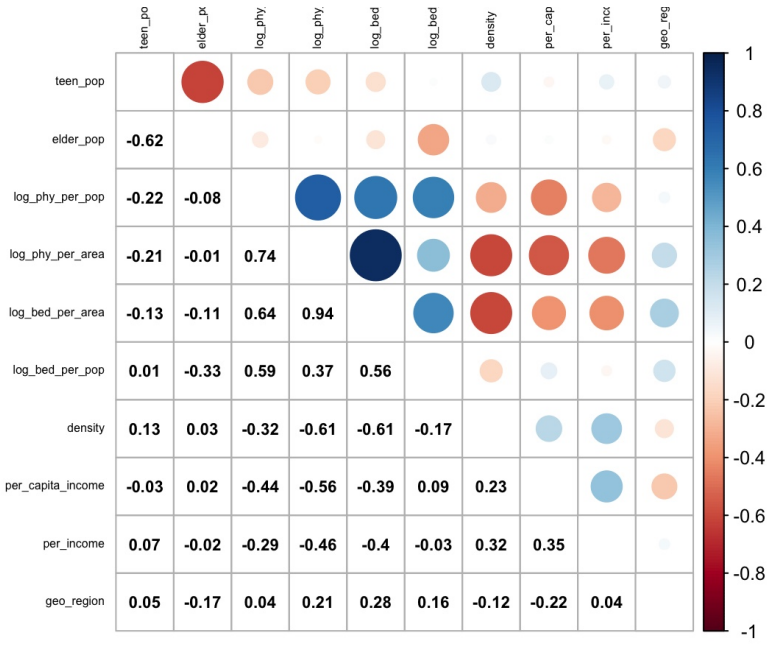


Figure 5: Correlation Heat Map

From the plot, we can get the following conclusions:

1. Teenager population rate is negatively related to elder population rate.
2. Log of population density is negatively related to log of hospital beds per area and log of hospital beds per population
3. Log of physicians per area and log of physicians per population are positively related to log of hospital beds per population and log of hospital beds per area.
4. Log of physicians per area and per population are negatively related to log of population density.

Therefore, we should pay special attention to those correlated predictors.

2.2.3 Interaction

We are also interested in the interaction between the variables. To explore the interaction between the variables in the logistic regressions, we plot the interaction plots to see if any interactions do exist between the variables.

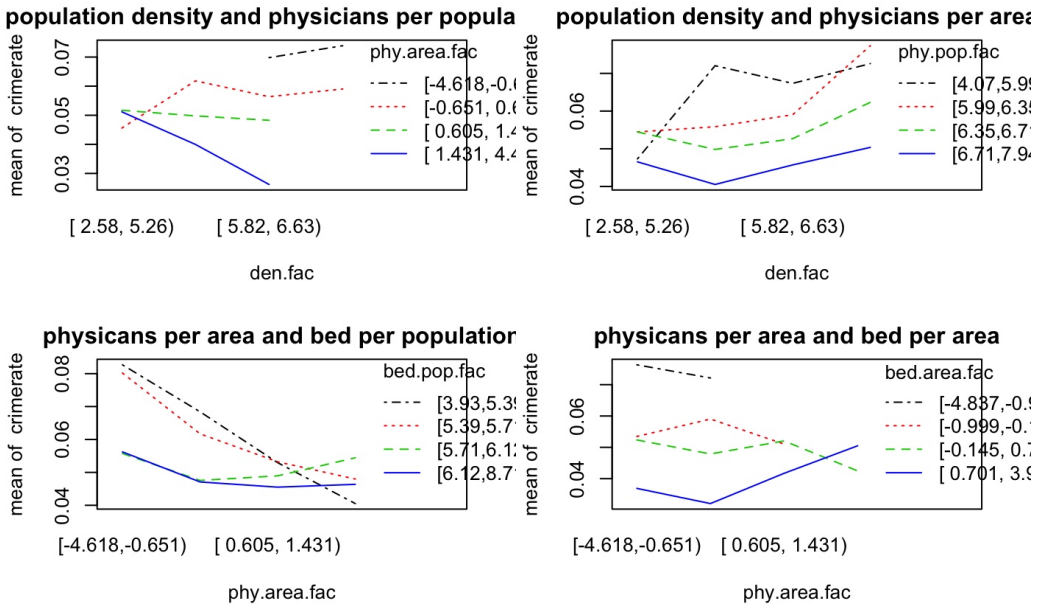


Figure 6: interaction plot-1

Since the four lines are not parallel to each other, we can conclude that there exist interactions between physicians and beds per area for four crime levels.

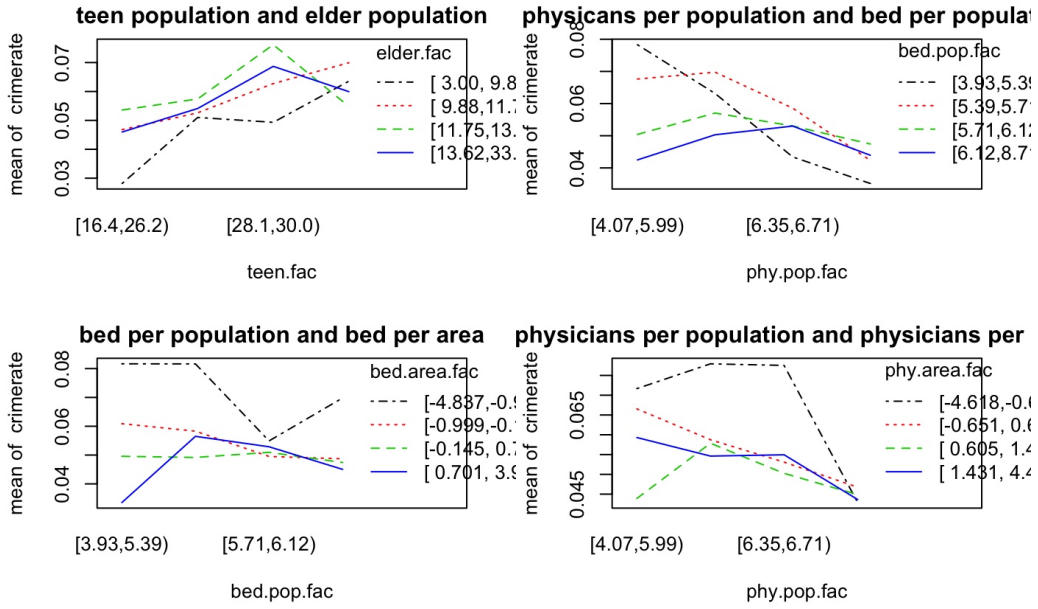


Figure 7: interaction plot-2

Similarly, from the above interaction plots, we can conclude that there are interactions between:

1. teen population and elder population.
2. physicians per area and physicians per
3. teen population and density
4. bed per population and below poverty
5. capital income and density
6. bed per population and geographical region

3 Analysis of results

3.1 Final model and interpretation

3.1.1 Final model

We chose level=1 as our base line and our final model is:

$$\begin{aligned} log(\frac{P(crime\_level = j)}{P(crime\_level = 1)}) =& \beta_{0j} + \beta_{1j} * teen\_pop + \beta_{2j} * log\_bed\_per\_pop) + \beta 3j * below\_poverty \\ &+ \beta 4j * log_density + \beta 5j * per\_capita\_income + \beta 6j * geo\_region \\ &+ \beta 7j * (teen\_pop * log_density)) \\ &+ \beta 8j * (log(bed\_per\_pop) * below\_poverty) \\ &+ \beta 9j * (log(bed\_per\_pop) * geo\_region) \\ &+ \beta 10j * (below\_poverty * per\_capita\_income) \\ &+ \beta 11j * (log(density) * per\_capita\_income) \end{aligned} \tag{1}$$

The following are our model predictors:

Variable Name	β_{j1}	β_{j2}	β_{j3}	p(1)	p(2)	p(3)
teen_pop	0.256	-0.272	-0.306	0	0	0
log_bed_per_pop	-4.007	-6.216	-3.846	0	0	0
below_poverty	-2.098	-3.224	-3.649	0	0	0
log_density	4.929	3.168	3.401	0	0	0
per_capita_income	0.001	0.001	0.001	0	0	0
geo_region	-0.838	-1.840	5.144	0	0	0
teen_pop*log_density	-0.038	0.069	0.082	0	0	0
log_bed_per_pop:below_poverty	0.370	0.560	0.585	0	0	0
log_bed_per_pop:geo_region	0.475	0.703	-0.391	0	0	0
below_poverty:per_capita_income	$2.08 * 10^{-5}$	$3.55 * 10^{-5}$	7.07*10−5	0	0	0
log_density:per_capita_income	-0.001	-0.002	-0.002	0	0	0

Table 3: model parameters

3.1.2 Interpretation

Variable Name	$e^{\beta_{j1}}$	$e^{\beta_{j2}}$	$e^{\beta_{j3}}$
teen_pop	1.291	0.762	0.736
log_bed_per_pop	0.018	0.002	0.021
below_poverty	0.123	0.040	0.026
log_density	138.190	23.768	30.008
per_capita_income	1.001	1.001	1.001
geo_region	0.432	0.159	171.438
teen_pop*log_density	0.962	1.072	1.085
log_bed_per_pop:below_poverty	1.447	1.749	1.795
log_bed_per_pop:geo_region	1.608	2.019	0.676
below_poverty:per_capita_income	1.000	1.000	1.000
log_density:per_capita_income	0.999	0.999	0.999

Table 4: exp(model parameters)

From the table above, we can easily see that (take the exponential of those values)

log_density*,*log_bed_per_pop:below_poverty*,*below_poverty

are the three most important factors which will influence the crime level of a region.

1. Interpretation of log_density

While holding other variables fixed, compared to the **crime level 1**,
if we increase **log_density** by 1:
crime level 2 will change by a multiplicative factor equal to $e^{4.929} = 138.19$,
crime level 3 will change by a multiplicative factor equal to $e^{3.168} = 23.76$
crime level 4 will change by a multiplicative factor equal to $e^{3.401} = 30$

2. Interpretation of log_bed_per_pop:below_poverty

While holding other variables fixed, compared to the **crime level 1**,
if we increase **log_bed_per_pop:below_poverty** by 1:
crime level 2 will change by a multiplicative factor equal to $e^{0.370} = 1.447$,
crime level 3 will change by a multiplicative factor equal to $e^{0.560} = 1.749$
crime level 4 will change by a multiplicative factor equal to $e^{0.585} = 1.795$

3. Interpretation of below_poverty

While holding other variables fixed, compared to the **crime level 1**,
if we increase **per_capita_income** by 1:
crime level 2 will change by a multiplicative factor equal to $e^{-2.098} = 0.123$,
crime level 3 will change by a multiplicative factor equal to $e^{-3.224} = 0.040$
crime level 4 will change by a multiplicative factor equal to $e^{-3.649} = 0.026$

3.2 Model selection

We tried the following different models, all fitted in logistis regression

Model.1	formula = as.factor(crime_level)$\sim$ $teen\_pop + elder\_pop + log\_phy\_per\_area + log\_bed\_per\_pop + bachelor + below\_poverty + high\_school + log\_density + per\_capita\_income + per\_income + geo\_region + unemployment$
Model.2	formula = as.factor(crime_level) $\sim$ $teen\_pop + log\_bed\_per\_pop + bachelor + below\_poverty + log\_density + per\_capita\_income + geo\_region$
Model.3	formula = as.factor(crime_level) $\sim$ $teen\_pop + log\_bed\_per\_pop + bachelor + below\_poverty + log\_density + per\_capita\_income + geo\_region + teen\_pop * log\_bed\_per\_pop + teen\_pop * bachelor + teen\_pop * below\_poverty + teen\_pop * log\_density + teen\_pop * per\_capita\_income + teen\_pop * geo\_region + log\_bed\_per\_pop * bachelor + log\_bed\_per\_pop * below\_poverty + log\_bed\_per\_pop * log\_density + log\_bed\_per\_pop * per\_capita\_income + log\_bed\_per\_pop * geo\_region + bachelor * below\_poverty + bachelor * log\_density + bachelor * per\_capita\_income + bachelor * geo\_region + below\_poverty * log\_density + below\_poverty * per\_capita\_income + below\_poverty * geo\_region + log\_density * per\_capita\_income + log\_density * geo\_region + per\_capita\_income * geo\_region)$
Model.4	formula = as.factor(crime_level)$\sim$ $teen\_pop + log\_bed\_per\_pop + below\_poverty + log\_density + per\_capita\_income + geo\_region + teen\_pop : log\_density + log\_bed\_per_pop : below\_poverty + log\_bed\_per\_pop : geo\_region + below\_poverty : per\_capita\_income + log\_density : per\_capita\_income$

Table 5: different models

To summarize:

For Model.1: we include all the variables in logistic regression
For Model.2: we used backward step-wise selection to choose variables that are significant to the model
For Model.3: we added all the interaction terms to model.2
For Model.4: we used backward step-wise selection to choose variables

Hence, we got our final model.

The AIC and BIC of all four models are:

	Model.1	Model.2	Model.3	Model.4
AIC	917.8823	899.0451	913.9511	857.2098
BIC	1077.267	997.1277	1269.5000	1004.334

Table 6: model evaluation

We can see from the above table that both the values of AIC and BIC improve a lot during our model selection process .

3.3 Diagnostics and model validation

3.3.1 Residual plots

After finding the final model, We would like to check the residual of it.

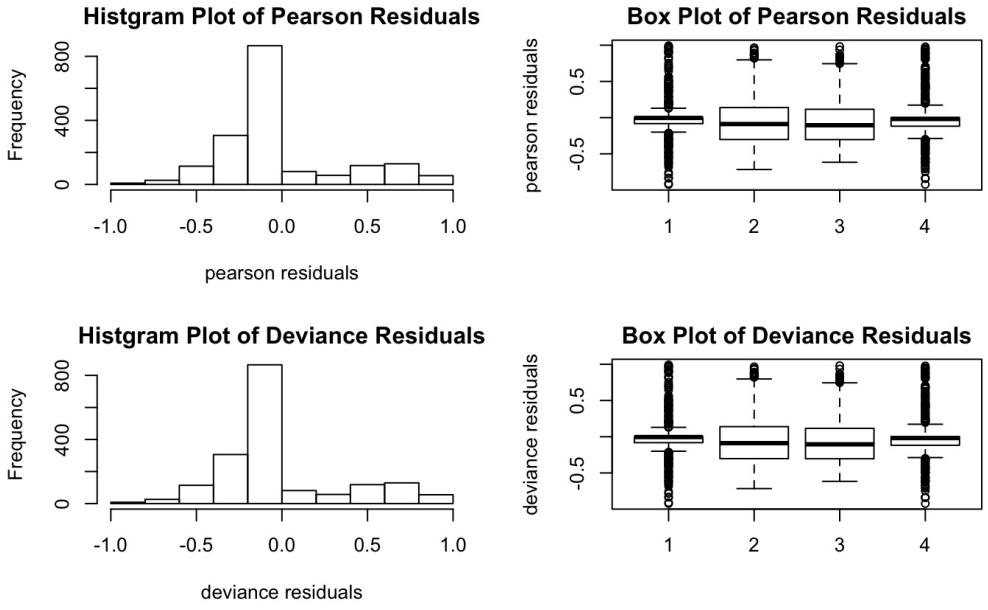


Figure 8: box plot of crime rate

In this step, we calculate the Pearson residual and deviance residual of our final model.

While both the Pearson residual and deviance residual show a little right-skewed, it is acceptable.

3.3.2 Model validation

In the model validation process, we use original data as our test dataset and we predict the new crime rate with the final model that we generate. Then, we compare the actual data with the prediction result to evaluate the performance of our final model.

	crime_level.1	crime_level.2	crime_level.3	crime_level.4
predict.1	90	20	8	1
predict.2	11	57	35	13
predict.2	6	23	47	15
predict.2	3	10	20	81

Table 7: Confusion Table

$$AccuracyRate = \frac{90 + 57 + 47 + 81}{440} = 62.5\% \tag{2}$$

According to the Confusion Table, the numbers on the diagonal position represent the correct predictions on that level. After summing the numbers in the diagonal positions, we calculate the accuracy of the prediction model on the original data set. In this case, the accuracy rate of our final model is 62.5%.

4 Conclusion

Variable Name	β_{j1}	β_{j2}	β_{j3}	$e^{\beta_{j1}}$	$e^{\beta_{j2}}$	$e^{\beta_{j3}}$
teen_pop	0.256	-0.272	-0.306	1.291	0.762	0.736
log_bed_per_pop	-4.007	-6.216	-3.846	0.018	0.002	0.021
below_poverty	-2.098	-3.224	-3.649	0.123	0.040	0.026
Plog_density	4.929	3.168	3.401	138.190	23.768	30.008
per_capita_income	0.001	0.001	0.001	1.001	1.001	1.001
geo_region	-0.838	-1.840	5.144	0.432	0.159	171.438
teen_pop*log_density	-0.038	0.069	0.082	0.962	1.072	1.085
bed_per_pop:poverty	0.370	0.560	0.585	1.447	1.749	1.795
bed_per_pop:region	0.475	0.703	-0.391	1.608	2.019	0.676
poverty:capital_income	$2.08 * 10^{-5}$	$3.55 * 10^{-5}$	$7.07 * 10-5$	1.000	1.000	1.000
density:capital_income	-0.001	-0.002	-0.002	0.999	0.999	0.999

Table 8: model parameters

For the interaction terms shown in the table above, we use the abbreviations to make sure our table is clean. bed_per_pop refers to log_bed_per_pop, poverty refers to below_poverty, region refers to geo:region, and capital_income refers to per_capital_income.

According to the final model we obtained, we got a few insights for the future policies adjustment.

First: the higher the population density, the higher the crime rate would like to be. So maybe one of the possible suggestion for government is to try to lower the population density of the area, develop the rural area.

Second: the personal income is also negatively correlated with the crime level. And the correlation between the percentage of the poverty population and the crime rate is positive. So the second step to lower the crime rate might be putting some effort to develop the economic of the local region and made some adjustment to the tax policy to help the people below the poverty line would be a good way to lower the crime rate.

Third: assigning more hospital bed would be helpful to lower the crime rate too.

5 Appendix

Final model

$$\begin{aligned} log(\frac{P(crimelevel=2)}{P(crimelevel=1)}) = &-15.987 + 0.256 * teenpop - 4.007 * log(bedperpop) - 2.097 * belowpoverty \\ &+ 4.929 * log(density) + 0.001 * percapitaincome - 0.838 * georegion \\ &- 0.038 * (teenpop * log(density)) \\ &+ 0.369 * (log(bedperpop) * belowpoverty) \\ &+ 0.475 * (log(bedperpop) * georegion) \\ &+ 2.088 * 10^{-5} * (belowpoverty * percapitaincome) \\ &- 1.432 * 10^{-4} * (log(density) * percapitaincome) \end{aligned} \tag{3}$$

$$\begin{aligned} log(\frac{P(crimelevel=3)}{P(crimelevel=1)}) = &2.092 - 0.272 * teenpop - 6.216 * log(bedperpop) - 3.224 * belowpoverty \\ &+ 3.168 * log(density) + 0.001 * percapitaincome - 1.840 * georegion \\ &+ 0.070 * (teenpop * log(density)) \\ &+ 0.559 * (log(bedperpop) * belowpoverty) \\ &+ 0.703 * (log(bedperpop) * georegion) \\ &+ 3.554 * 10^{-5} * (belowpoverty * percapitaincome) \\ &- 1.828 * 10^{-4} * (log(density) * percapitaincome)) \end{aligned} \tag{4}$$

$$\begin{aligned} log(\frac{P(crimelevel=4)}{P(crimelevel=1)}) = &-17.540 - 0.306 * teenpop - 3.846 * log(bedperpop) - 3.649 * belowpoverty \\ &+ 3.401 * log(density) + 0.001 * percapitaincome + 5.144 * georegion \\ &+ 0.082 * (teenpop * log(density)) \\ &+ 0.585 * (log(bedperpop) * belowpoverty) \\ &- 0.391 * (log(bedperpop) * georegion) \\ &+ 7.036 * 10^{-5} * (belowpoverty * percapitaincome) \\ &- 1.904 * 10^{-4} * (log(density) * percapitaincome)) \end{aligned} \tag{5}$$