Report of MA678 Midterm Project

Keliang Xu

12/10/2021

Abstract

Crime incident reports are provided by Boston Police Department (BPD) to document the initial details surrounding an incident includes times, locations, and descriptions of crimes. According to different crime rate in district, my question is: What does the crime rate in a certain district associated with? To dig deeper into this issue, I use one categorical factor and some demographic factors related to crime rate and build multilevel model. The model shows that income and labor rate have negative impact and poverty has positive impact on crime rate in Boston and is different between district. This report are consisted 5 main parts: Introduction, Method, Result and Discussion.

Introduction

Crime incident reports are provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. It is containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Because the data is the information of each case, after data integration, the number of each type of crime in each district at each time is counted. In this dataset, I couldn't see the relevant factors, so I introduced demographic data in Boston. Convert the number of crimes into crime rates, and build a model with the factor in demographic. Among a lot of demographic data, I chose three representative data, income, poverty rate and labor rate.

Therefore, I use a multilevel model to see how these factors affect the crime rate. Before that, I will clean up the data and combine demographic data in Boston.

Method

Data Cleaning and Processing

The main data set I found pobulished on Kaggle: Crimes in Boston. And I combine the main data set with demographic data for Boston's Neighborhoods on Demographic Data for Boston's Neighborhoods.

Firstly, because raw data recorded case by case, I need to count them according to a certain rule. I chose district, month and type of crime(UCR_PART). Plus, the time period of time is 2015.7-2018.8 and I use the average crime number of each month. Secondly, I filtered and removed empty, `other` and obviously unreasonable data in new data set. Thirdly, by marking cases on the map, determine which Neighborhoods each district represents. Finally, I combine the cleaned data with demographic data for Boston's Neighborhoods so I get a brand new data set for next-step modeling.

In order to facilitate the modeling below, I performed some transformations on `income` and `crime_rate`. The final tidy data set has 432 rows and 7 rows which contain all the data and variables I use in this report.

Here are explanations of all columns I used:

column names	explanation
UCR_PART	UCR crime categories
Month	Month the crime occurred
District	The district in Boston
log_Income	The log of Per Capita Income
Labor_rate	The rate of labor force participation among 16+
Poverty_rate	The rate of poverty
Crime_rate	Crime rate per 100 people

Exploratory Data Analysis

In the tidy data set, there are three continuous variables `log(income)`, `labor_rate`, and `poverty_rate`. I make some plots in order to clearly show the distribution of continuous variables and the correlation between variables and `crime_rate`.

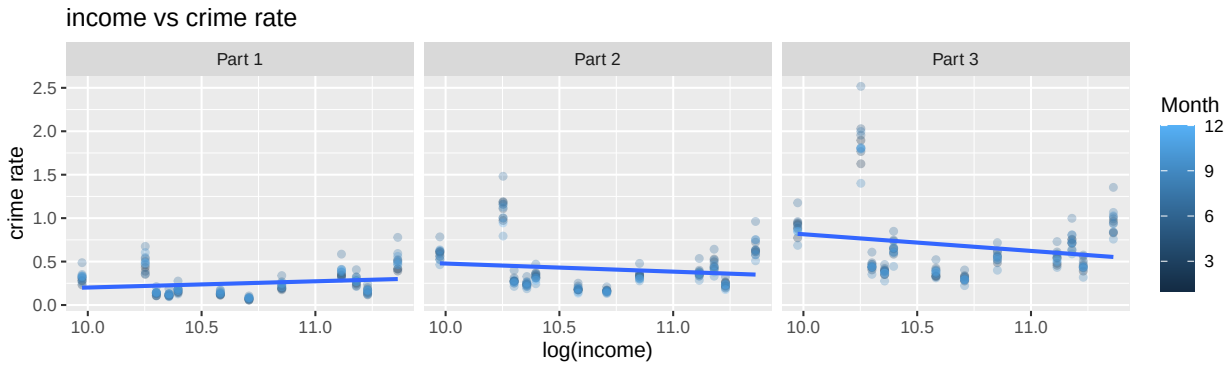


Figure 1: Correlation between log(income) and crime_rate(per 100 people)

Figure 1 shows the relationship between income and crime rate in district. The three plots represent three types of crimes (UCR_PART). The slopes of Part 2 and Part 3 plots are negative, indicating negative correlations, but the slope of Part 1 plot is positive but very close to zero. In order to continue to explore this difference, the crime type `UCR_PART` will be added to the model as a categorical variable later.

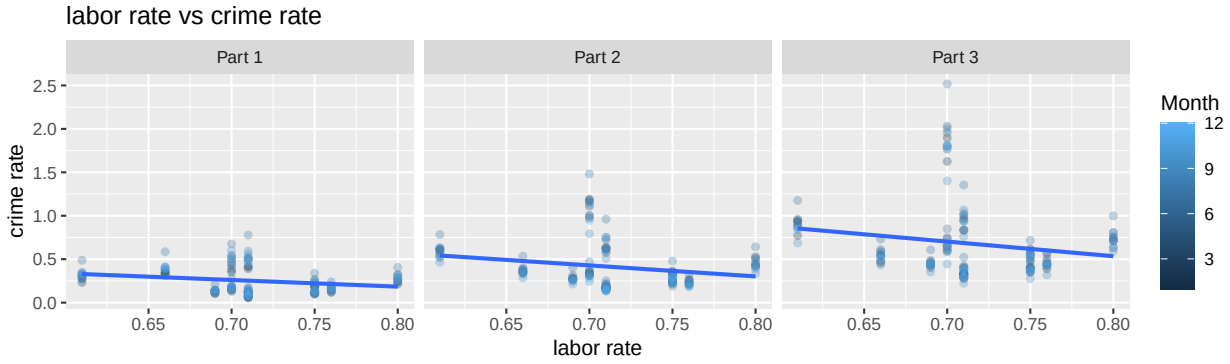


Figure 2: Correlation between labor rate and crime_rate(per 100 people)

Figure 2 shows the relationship between labor rate among 16+ and crime rate in district. It can be seen from the slope that all three plots show the negative correlation between labor rate and crime rate. It is worth mentioning that the slope of the part 3 plot is the steepest, which is consistent with the one shown in figure 1.

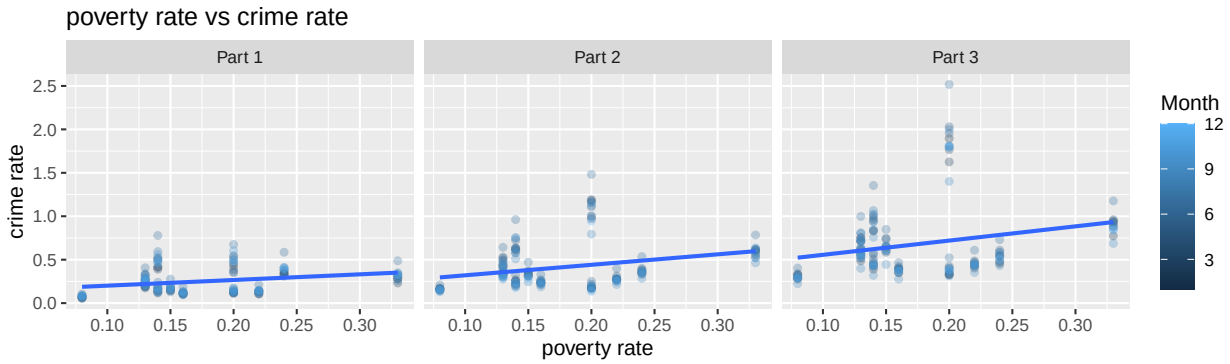


Figure 3: Correlation between poverty rate and crime_rate(per 100 people)

Different from the previous figures, the slopes of these three plots are positive. It shows the positive correlation between poverty rate and crime rate. At the same time, the slope of Part 3 is still the steepest one. It is guessed that the type of crime Part 3 is more related to the demographic data.

Model Fitting

In order to consider different district, I use multilevel model to fit the data. It is clear show the different correlation with three continues variables log(income), labor rate, poverty rate with different `UCR_PART`. So I will put `UCR_PART` as categorical variable in the model. Below is the function:

```
model<-lmer(Crime_rate~UCR_PART+log_Income+Labor_rate+Poverty_rate+(UCR_PART|District)
+(1+log_Income|District)+(1+Labor_rate|District)+(1+Poverty_rate|District),newtidydata)
```

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-2.16	0.72	9.72	-3.013	0.013455 *
UCR_PARTPart 2	0.16	0.05	11.01	3.528	0.004729 **
UCR_PARTPart 3	0.43	0.09	11.00	4.608	0.000756 ***
log_Income	0.02	0.07	13.31	4.278	0.000853 ***
Labor_rate	-0.46	0.63	3.64	-0.741	0.503423
Poverty_rate	0.92	0.40	9.77	2.304	0.044542 *

Result

Model Coefficients

As an example, the formula in A1 district:

$$crime\_rate = -2.11 + 0.18 \cdot UCR\_PART_{Part2} + 0.48 \cdot UCR\_PART_{Part3} + 0.34 \cdot log(income) + -0.37 \cdot Labor\_rate + 0.92 \cdot Poverty\_rate$$

The coefficients of `UCR_PART` are also in line with the previous observations. Part 3 has the steepest slope, and its coefficient is positive and greater than that of Part 2. In the case where the three continuous variables are the same, when Part 2 replaces Part 1, the predicted difference in crime rate decreases by 2.11%. Next, the positive and negative coefficients of the three continuous variables in the formula are consistent with EDA plots. For each 1% increase in log(income), the predicted difference in crime rate increases by 0.34%. And the same for number of labor rate and poverty rate.

For different district, the influence of UCR_PART and two continuous are always not the same, while the magnitude of the difference of poverty rate is 10^{-7}. The small differences in this continuous variable may be due to the relatively small effect of different regions on the correlation between poverty and crime rates, and the fact that the other variables do not fluctuate much.

$Month	(Intercept)	Part 2	Part 3	log_Income	Labor_rate	Poverty_rate
A1	-2.11	0.18	0.47	0.34	-0.37	0.92
A7	-2.37	0.09	0.28	0.15	-0.70	0.92
A15	-2.34	0.10	0.28	0.25	-0.04	0.92
B2	-1.94	0.24	0.58	0.22	-0.57	0.92

Model Validation

For each coefficient in the function, I think it is reasonable. Among them, the intercept is negative because Part 1 crime rate is lower, and log income multiplied by the coefficients can make the whole formula results in positive. the coefficients before Part 2,3 are positive, which means that the crime rate of these two categories is higher than the first category. At the same time, the lower the labor rate, and the higher the poverty the less likely to commit a crime is easy to make sense. The coefficient of log(income) is positive but tends to 0, and can reflect the different slopes in figure 1.

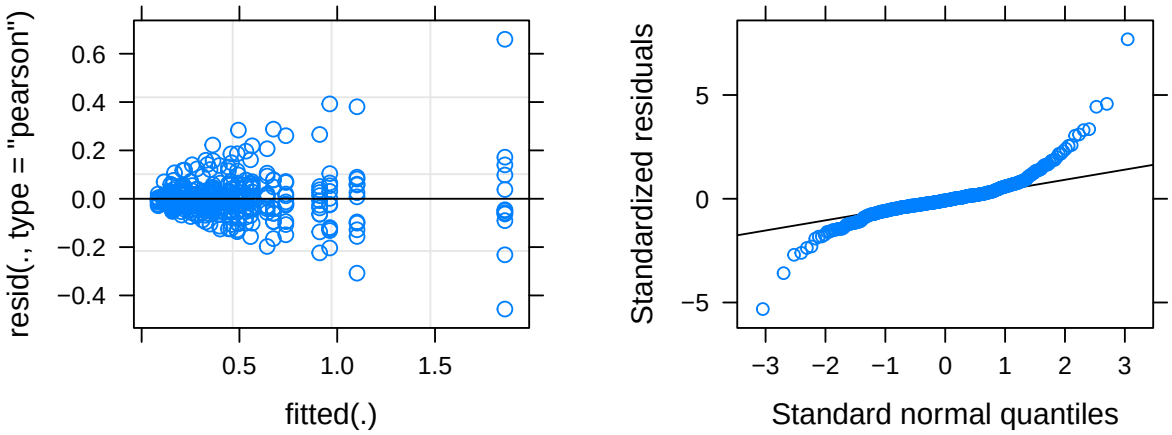


Figure 4: Residual plot and Q-Q plot.

Residual plot in figure 4 shows the mean of residuals is almost 0, but there are a few points that deviate more from 0. Similarly, the Q-Q plot shows well alignment to the the line with a few points at the top and bottom slightly offset. Figure 5 shows that most of points are within 95% confidence interval.

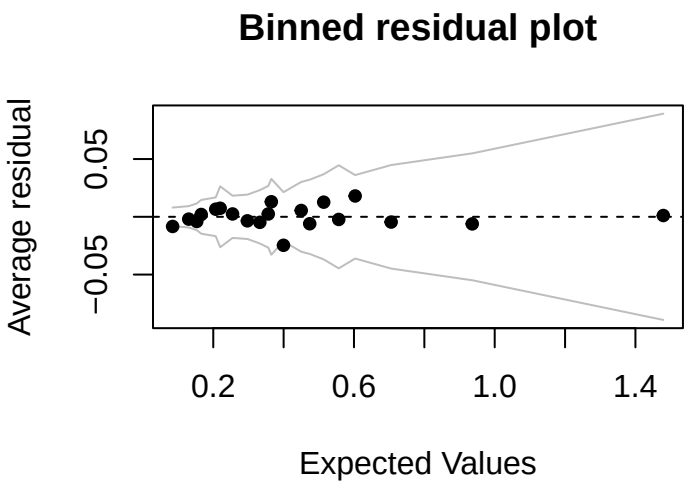


Figure 5: Binned Residual Plot.

Discussion

Conclusion

Through the above analysis and model validation, I think the model is reasonable in some extents. The model shows that income and labor rate have negative impact and poverty has positive impact on crime rate in district of Boston and is slightly different between district. The different estimates of these predictors are also convincing on the conditions of the different districts.

Limitation and Next step

1. The amount of data is very small. Do not look at the initial dataset has a total of 300,000+ pieces of data, but after sorting statistics, the amount of data left that can be used for modeling plummets. Moving forward, therefore, would then require expanding the regional selection/district to extend the total sample of crime studies from Boston further out to Massachusetts or the nation.

2. Limitation for the selection of demographic data. There are 23 categories of total demographic data classification, while the selection is mainly based on individual subjective wishes. There may be demographic data categories that are more likely to influence crime rates that were not detected. If the next step is carried out, all demographic data can be modeled and filtered.

3. Rougher data processing. The first one is the correspondence of district. There are inevitable errors in data collection, but I did not sort them according to latitude and longitude in data processing. The second is that population data is usually in years, so it seems difficult to use months as a reference and to expand the data more. In other words, if there is no way to get the population data by month, the model can only be structured on a year basis, which is not as accurate. These are now more difficult problems to deal with.

Appendix

Citation

Regression with Categorical Variables: Dummy Coding Essentials in R http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

R Bootcamp: Introduction to Multilevel Model and Interactions https://quantdev.ssri.psu.edu/tutorials/r-bootcamp-introduction-multilevel-model-and-interactions

Comparison table of DISTRICT

DISTRICT	Neighborhoods Name
A1	Beacon Hill,Downtown,North End,West End
A7	East Boston
A15	Charlestown
B2	Mission Hill,Roxbury
B3	Mattapan
C6	South Boston,South Boston Waterfront
C11	Dorchester
D4	Back Bay,Fenway,South End
D14	Allston,Brighton
E5	Roslindale,West Roxbury
E13	Jamaica Plain
E18	Hyde Park

More EDA

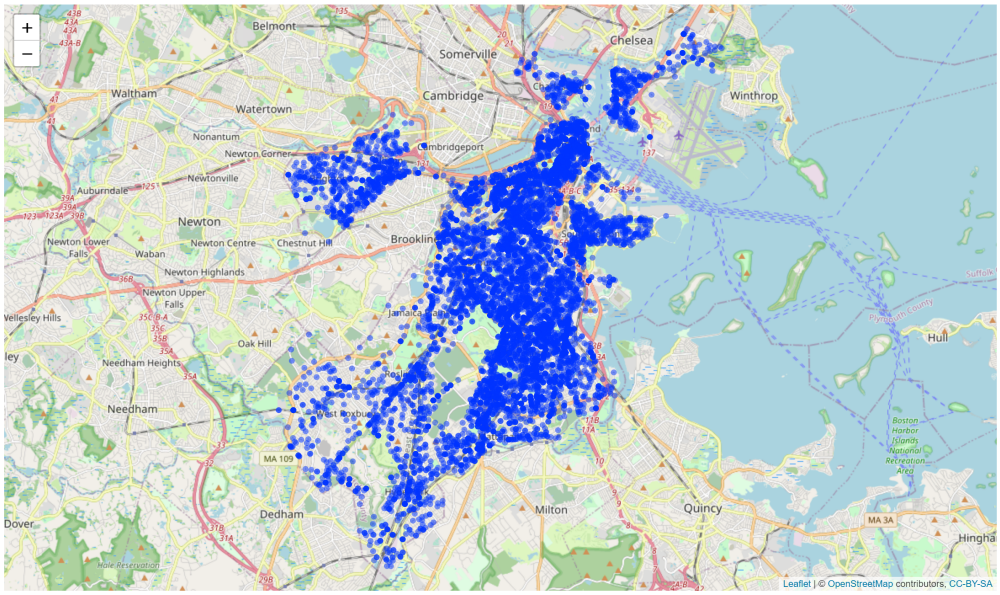
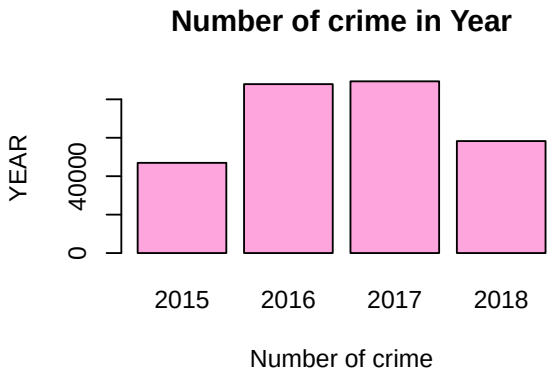
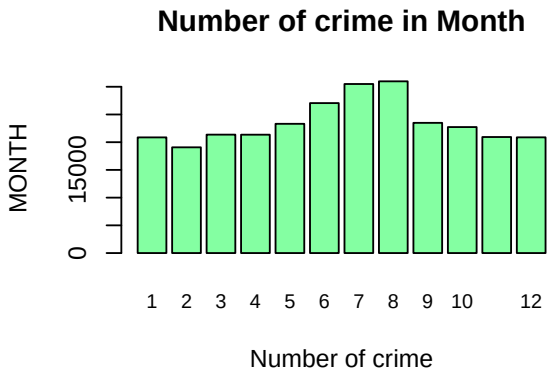
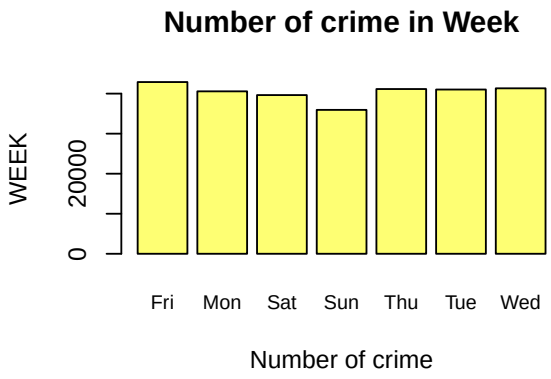
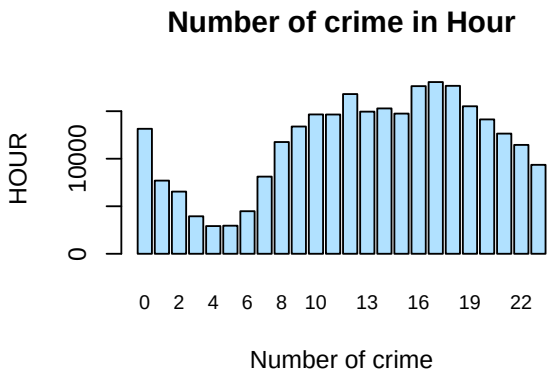


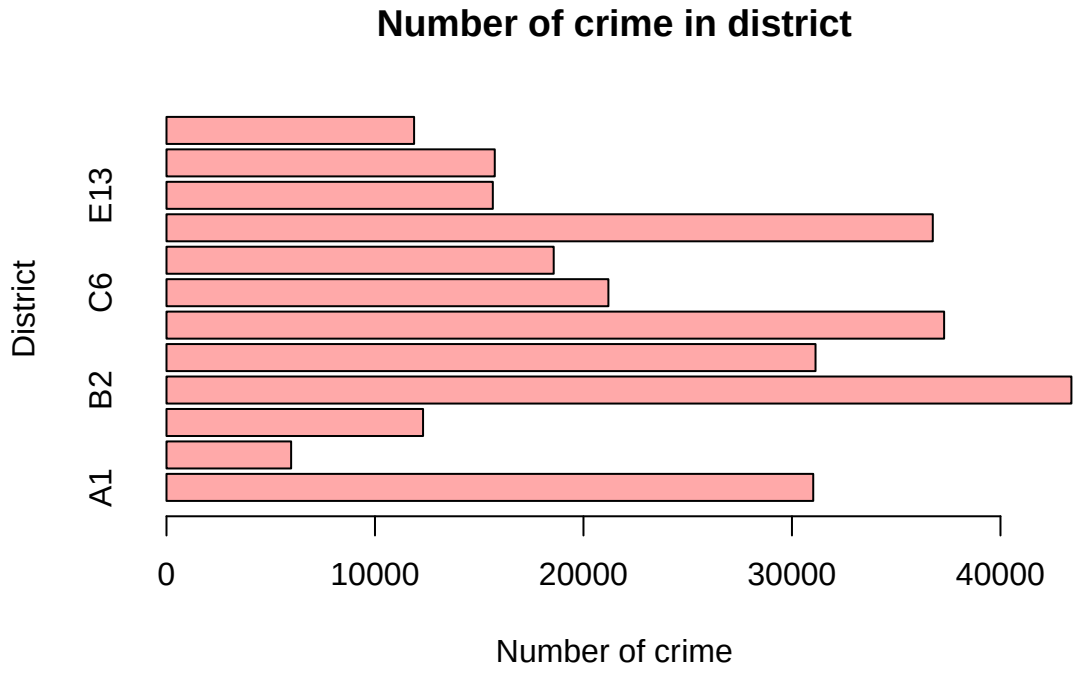
Figure 6: Map of crime data.


Number of crime in Year
YEAR
0
40000
2015
2016
2017
2018
Number of crime


Number of crime in Month
MONTH
0
15000
1
2
3
4
5
6
7
8
9
10
12
Number of crime


Number of crime in Week
WEEK
0
20000
Fri
Mon
Sat
Sun
Thu
Tue
Wed
Number of crime


Number of crime in Hour
HOUR
0
10000
0
2
4
6
8
10
13
16
19
22
Number of crime


Number of crime in district
District
E13
C6
B2
A1
0
10000
20000
30000
40000
Number of crime

Full Results Random effects of model

```
## $District
##      (Intercept) UCR_PARTPart 2 UCR_PARTPart 3  (Intercept)    log_Income
## A1   0.012406826     0.01733018     0.03645910 -1.024726459  0.0995473125
## A15 -0.052856015    -0.07371928    -0.15004316  0.908976537 -0.0883027569
## A7  -0.047372623    -0.06627623    -0.14414849 -0.123383429  0.0119861146
## B2   0.053917122     0.07517185     0.15175872  0.173016221 -0.0168077048
## B3   0.322666955     0.45049171     0.93778439 -0.006234908  0.0006056918
## C11 -0.043455780    -0.06071875    -0.12856853  0.006337895 -0.0006156965
## C6   0.008268703     0.01154017     0.02383385  0.027521815 -0.0026736137
## D14 -0.074378467    -0.10374035    -0.21129727  0.546054433 -0.0530465968
## D4  -0.091181299    -0.12721394    -0.26079927 -0.315276976  0.0306276621
## E13 -0.032058114    -0.04476753    -0.09362225 -0.353053155  0.0342974385
## E18  0.013868877     0.01941508     0.04276780  0.023545079 -0.0022872927
## E5  -0.069826185    -0.09751291    -0.20412489  0.137222946 -0.0133305580
##      (Intercept)   Labor_rate   (Intercept)  Poverty_rate
## A1  -0.061600214  0.090806936  9.724211e-09 -3.853777e-08
## A15  0.160927018 -0.237227901 -9.844545e-09  3.901466e-08
## A7  -0.284653695  0.419617533  1.629844e-08 -6.455195e-08
## B2   0.075238899 -0.110912177 -3.833394e-09  1.529721e-08
## B3   0.006634154 -0.009779629 -7.104655e-10  2.806373e-09
## C11  0.020525689 -0.030257605 -2.513192e-09  9.888974e-09
## C6   0.007667152 -0.011302405 -3.429122e-10  1.359304e-09
## D14  0.121522123 -0.179139827 -8.900290e-09  3.515657e-08
## D4   0.014317678 -0.021106167  4.150153e-10 -1.605518e-09
## E13 -0.091914401  0.135494094  6.980151e-09 -2.766931e-08
## E18  0.010076074 -0.014853477 -2.118010e-09  8.391489e-09
## E5   0.021259523 -0.031339374 -5.155009e-09  2.044996e-08
##
## with conditional variances for "District"
```

Fixed effects of model

```
##    (Intercept) UCR_PARTPart 2 UCR_PARTPart 3     log_Income     Labor_rate
##     -2.1553196      0.1617797      0.4306356      0.2402231     -0.4645556
##   Poverty_rate
##      0.9203767
```

Coefficients of model

```
## $District
##     (Intercept) UCR_PARTPart 2 UCR_PARTPart 3 log_Income  Labor_rate
## A1   -2.1056923     0.17910989      0.4670947  0.3397704 -0.37374867
## A15  -2.3667436     0.08806042      0.2805924  0.1519203 -0.70178350
## A7   -2.3448100     0.09550348      0.2864871  0.2522092 -0.04493807
## B2   -1.9396511     0.23695156      0.5823943  0.2234154 -0.57546778
## B3   -0.8646517     0.61227141      1.3684199  0.2408288 -0.47433523
## C11  -2.3291427     0.10106095      0.3020670  0.2396074 -0.49481321
## C6   -2.1222447     0.17331988      0.4544694  0.2375495 -0.47585801
## D14  -2.4528334     0.05803935      0.2193383  0.1871765 -0.64369543
## D4   -2.5200447     0.03456577      0.1698363  0.2708507 -0.48566177
## E13  -2.2835520     0.11701218      0.3370133  0.2745205 -0.32906151
```

```
## E18  -2.0998440     0.18119478      0.4734034  0.2379358 -0.47940908
## E5   -2.4346243     0.06426680      0.2265107  0.2268925 -0.49589498
##     Poverty_rate
## A1     0.9203766
## A15    0.9203767
## A7     0.9203766
## B2     0.9203767
## B3     0.9203767
## C11    0.9203767
## C6     0.9203767
## D14    0.9203767
## D4     0.9203767
## E13    0.9203766
## E18    0.9203767
## E5     0.9203767
##
## attr(,"class")
## [1] "coef.mer"
```

More residual plots

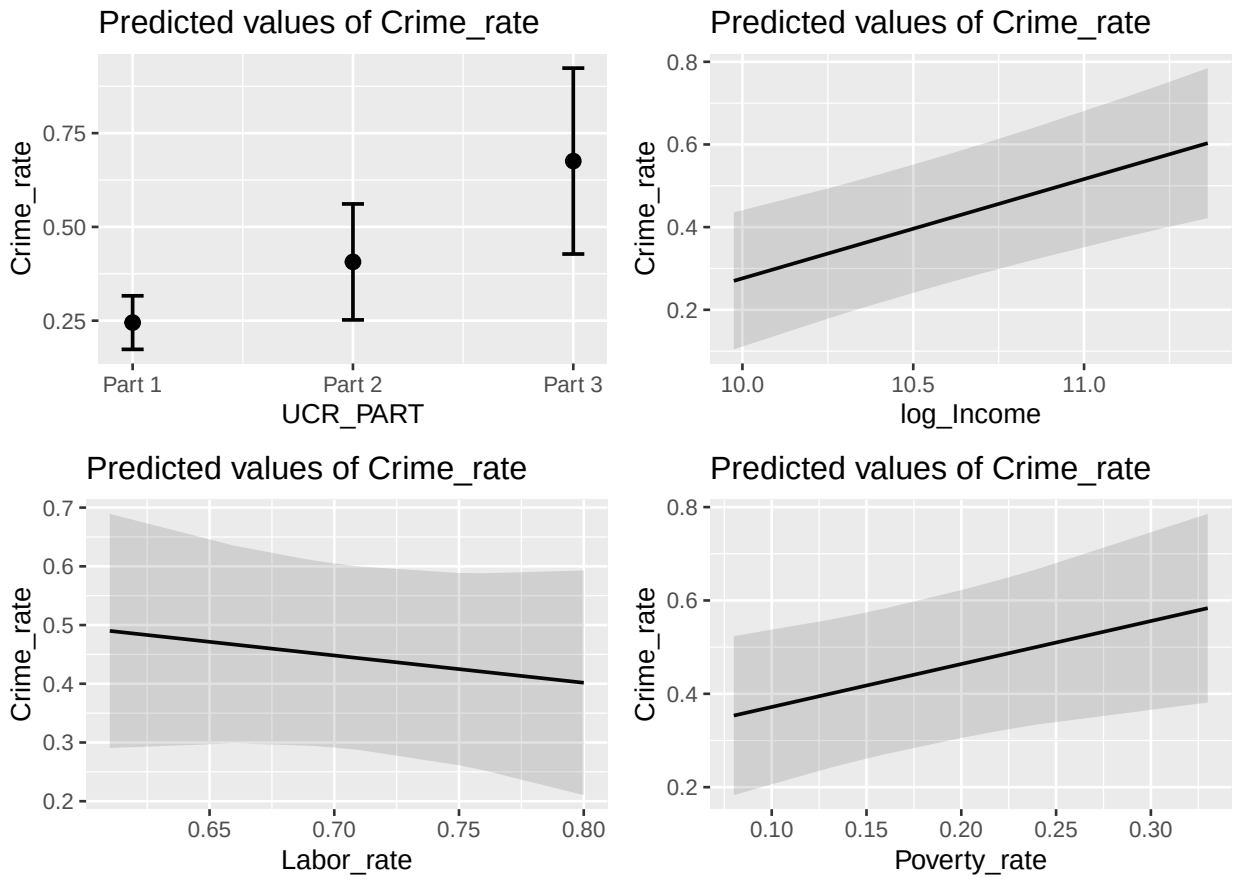


Figure 7: Predicted values (marginal effects).

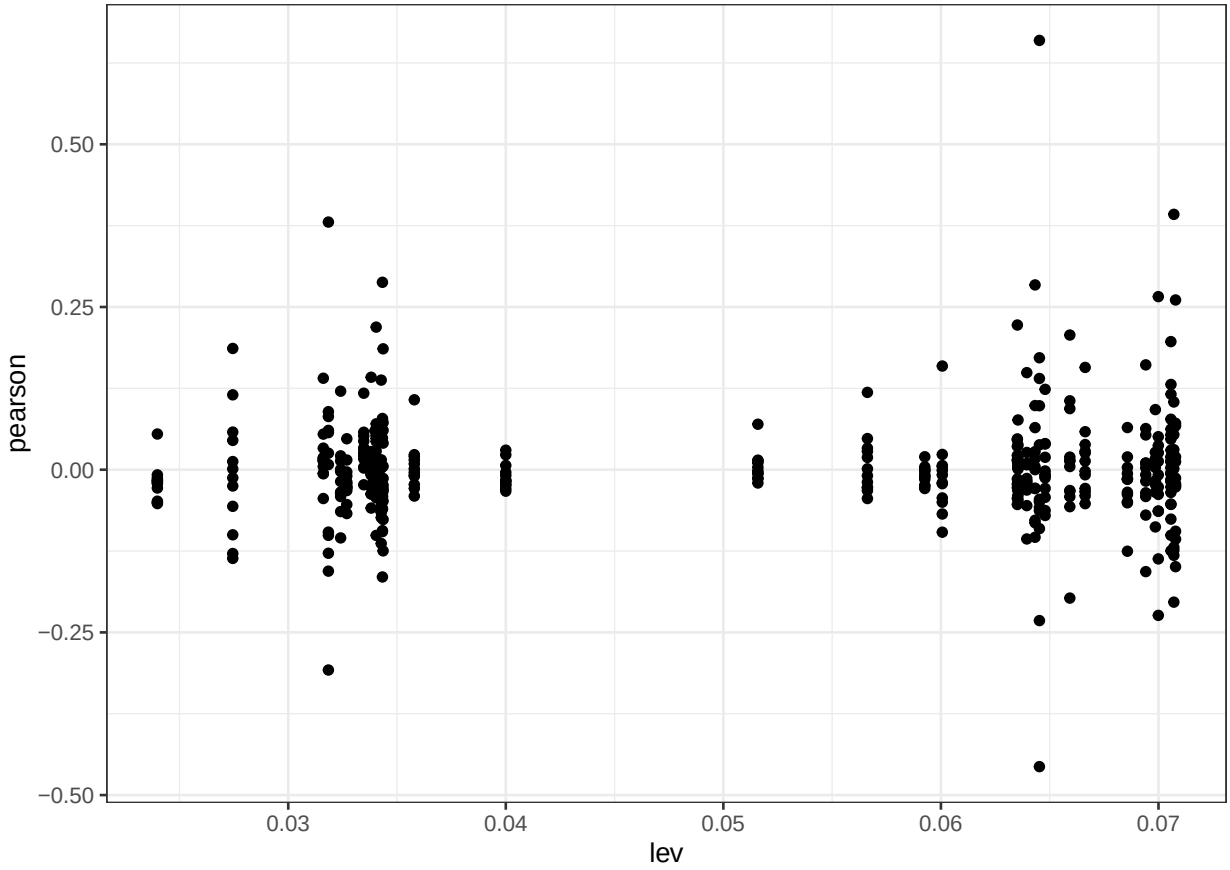


Figure 8: Residuals vs Leverage.