BIST8130 - Final Project Report

Mengfan Luo (ml4701) Yushan Wang (yw3772)
Jing Lyu (jl6049) Yiqun Jin (yj2686) Mingkuan Xu (mx2262)

2021/12/11

Abstract

In this project, we aim to build regression models based on a set of demographic variables to estimate county-level crime rates. After an exploratory analysis of the variables on their distributions and correlations, we derived more meaningful variables by manipulating the existing ones, removed outlier values, and implemented several variable selection methods. Using the selected variables, we trained a linear regression model, elaborated on the model by adding several interactive terms, and did cross-validations. All 3 resulting models have achieved a good estimation of the training set. We selected the third model (adjusted R-square: 0.542, RMSE: 16.46, RMSPE: 11.90) as the final model, as it has the best prediction on testing set. Further studies can be done on correcting the dataset using external data sources, as well as using more sophisticated non-linear models.

Introduction

Over the last three decades, crime has become a major public concern in the US arousing massive political discussion and public expenditure[1]. Crime rates in major cities experienced a general rise from the 1960s to 1990s, with two peaks observed in 1980 and in early 1990s[2]. Despite extensive attention across the nation, factors influencing crime trends were not yet made clear[1]. In this project, we examined crime rate and potential factors that affect the crime rate in "County

Demoraphic Information" (CDI), and constructed multiple linear regression model to predict crime rate.

Methods

Data Description

We analyzed data from the "County Demographic Information" (CDI) data set, which contains characteristics of 440 counties in the United States collected from 1990-1992. The primary goal of this investigation is to develop insight relevant to predicting the crime rate in counties.

Data Preprocessing Transform vairables in order to extract interpretable information.

Exploratory Analysis Calculate the pairwise correlations between variables and ist all correlations between the crime rate (our interest) and other variables.

Statistical Methods

Training/Testing Set Split: Randomly split the dataset into training (90%) and testing sets (10%) to avoid overfitting.

Remove Outliers and High Leverage Points: Use percentile to detect potential outliers and high leverage points. Remove rows containing the smallest and largest 0.2% for each variable given the size of the dataset.

Variable Selection: Select key variables using stepwise regression and criteria based procedure and conduct interaction analysis,

Model Construction: Build linear regression models using selected variables, interaction terms (p-value < 0.2). Attempt model transformations.

Cross Validation: Use 5-fold cross validation on each model to estimate model performances.

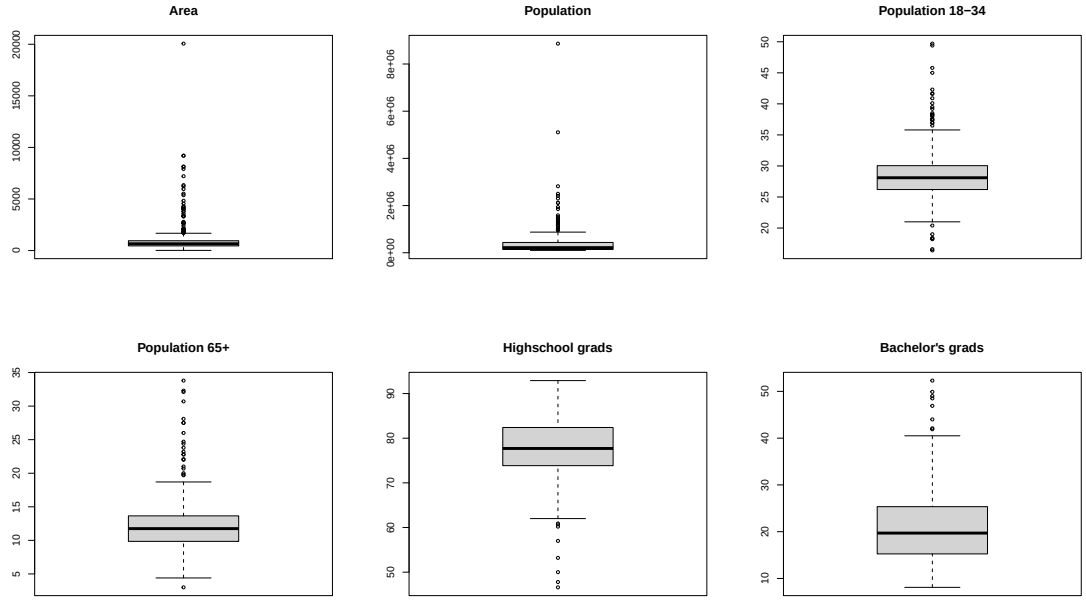
Model Assessment: Assess the models and choose a final model used for future prediction. Compare the models based on three criteria: the R-square values, the root mean square error (RMSE), and the root mean square prediction error (RMSPE).

- R-square value represents the proportion of the variance that can be explained by the regression model.
- RMSE measures the differences between the actual values and the predicted values in the training dataset.
- RMSPE estimates the prediction errors on new data outside the training dataset.

Results

Descriptive Analysis

After importing the csv file containing the County Demographic Information (CDI) data, we noticed that crimes, physicians, and hospital beds are given as numbers, while other info are given as proportions. We therefore computed the number of crimes, physicians, and hospital beds per 1000 people. Population density could be a key factor to crime rate. Thus, we also derived a new variable, `density`, which is population divided by area.



After drawing boxplots to show the distribution of the variables, we identified several extreme values in each of the variables. These values can be treated as potential outliers to be removed in further analysis. For example, the distribution of crime rate:

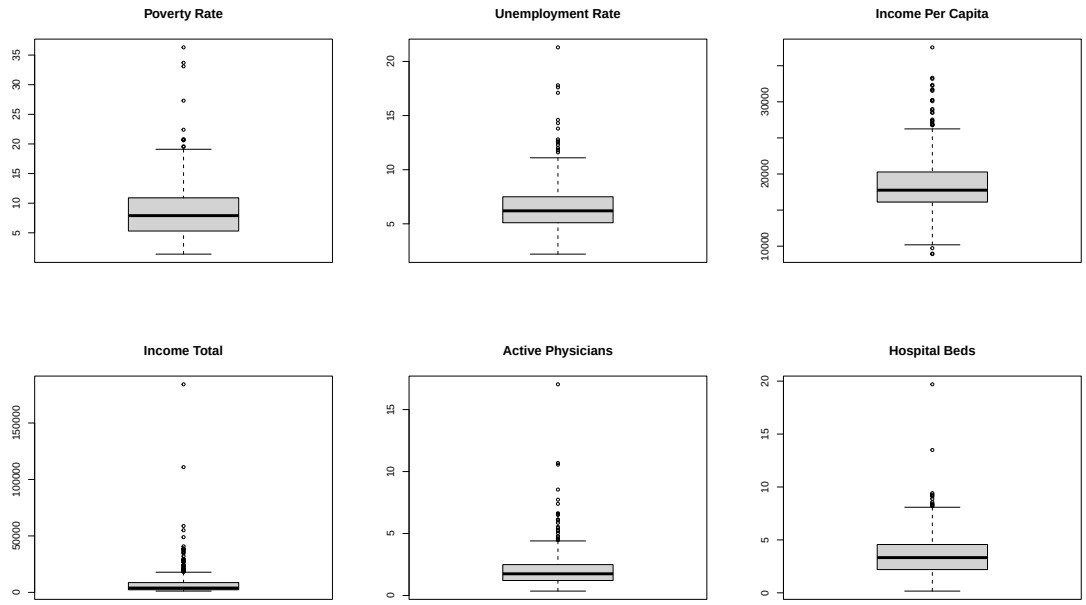


Figure 1: boxplot of continuous variables distribution

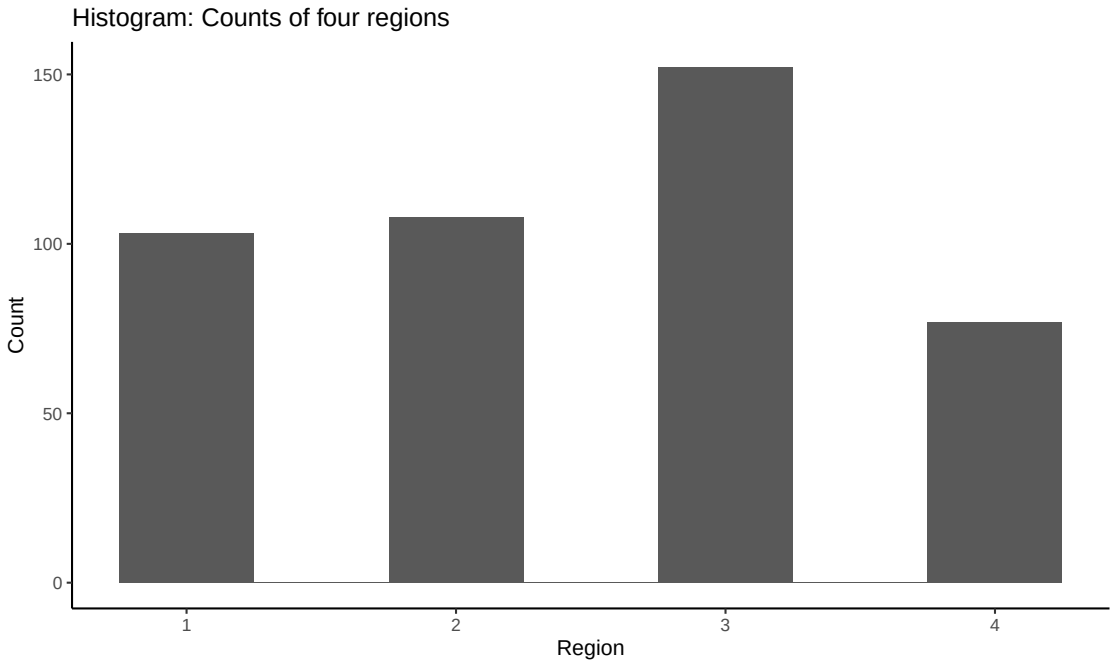


Figure 2: Histogram of catagorical variable:region distribution

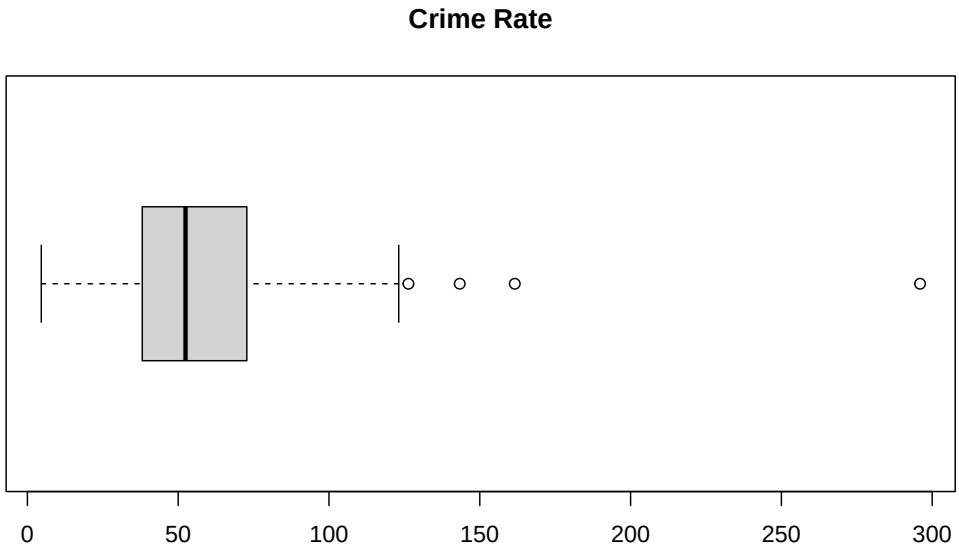


Figure 3: boxplot of dependent variable: crime rate

From the pairwise plots, we did not recognize any relationship between each variable. After taking a closer look of each variable, we calculated pairwise correlations and listed all the correlations between the crime rate (our interest) and other variables. Correlation analysis suggested that the derived variable, density, is a more meaningful variable compared to area and population, with a stronger association to the crime rate.

After a preliminary analysis of the data, we identified several variables that might be relevant to the crime rate as listed:

Table 1: Potential Variables Relevant to the Crime Rate

Variable	Meaning
area	Land area
density	Population Density
pop	Esimate 1990 population
pop18	Percent of population aged 18-34
pop65	Percent of population aged 65+

Variable	Meaning
docs_rate_1000	Number of active physicians per 1000 people
beds_rate_1000	Number of hospital beds per 1000 people
crime_rate_1000	Number of serious crimes per 1000 people
hsgrad	Percent high school graduates
bagrad	Percent bachelor's degrees
poverty	Percent below poverty level
unemp	Percent unemployment
pcincome	Per capita income
totalinc	Total personal income
region	Geographic region

Traning/Testing Split

We randomly sampled 10% from the dataset as a testing set (n = 44) and put the rest into a training set (n = 396).

Data Cleaning

We removed 19 rows with outliers and high leverage points from a total of 396 training data.

Variables Selection

Based on stepwise procedure, we selected the following variables:

Table 2: Vairable selected from stepwise regression

backward	stepwise
pop	pop
pop18	pop18

backward	stepwise
bagrad	bagrad
poverty	poverty
pcincome	pcincome
totalinc	totalinc
region2	region2
region3	region3
region4	region4
beds_rate_1000	beds_rate_1000

According to the output of criteria based procedure, we determined that the number of variables should be above 12 because $C_p \leq p$. Based on this analysis, we found that `unemp` and `density` could also be selected.

In addition, We removed `totalinc`, because it can be replaced with totalinc = pcincome * pop.

Interaction Analysis

Interaction term 1: `poverty` + `income`

According to the Census Bureau, the number of people below the official government poverty level was 33.6 million in 1990, representing 13.5 percent of the national population [4]. Thus, we used this criteria to divide `poverty` into two categories: higher than national poverty rate and lower than national poverty rate. In figure 6, we observed an intersection of the two lines, suggesting that the association between crime rate and per capita income is modified by poverty status.

Interaction term 2: `pcincome` + `bagrad`

According to the Census Bureau, the percentage of people 25 years old or older with bachelor's degrees was 20.8% in 1990 [4]. Thus, we used this criteria to divide `bagrad` into two categories: higher than national `bagrad` and lower than national `bargrad`. In figure 7, we observed that the association between crime rate and per capita income is modified by percent bachelor's degrees status.

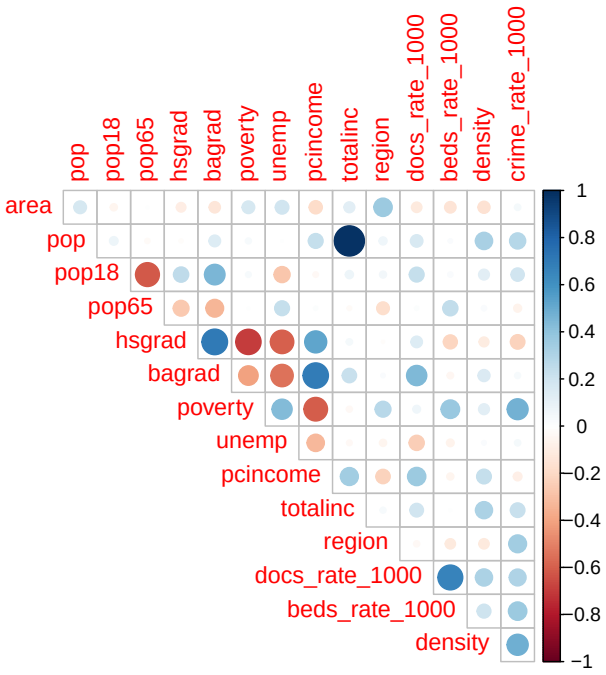


Figure 4: Correlation heatmap

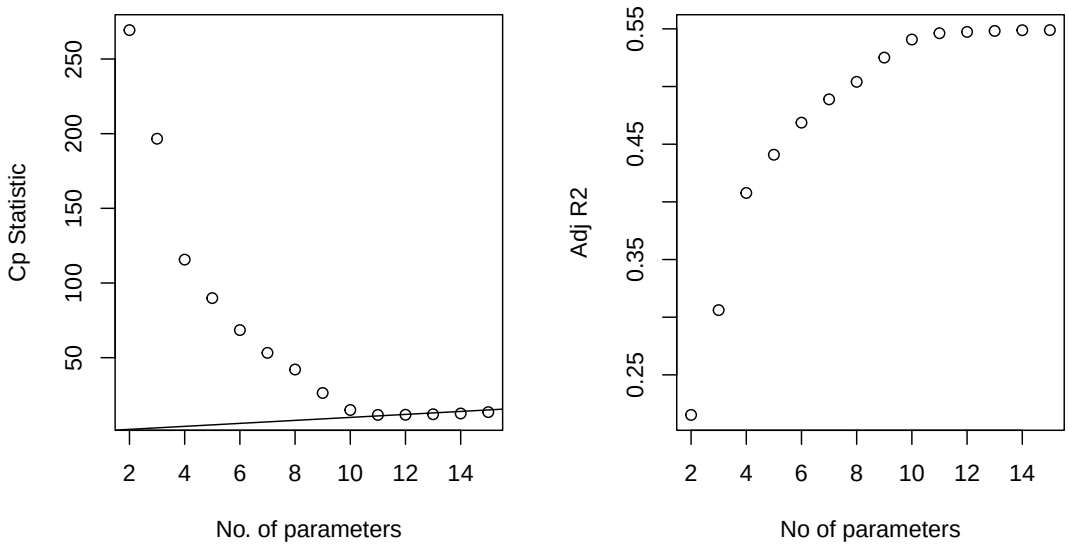


Figure 5: Subset selection for best parameter numbers

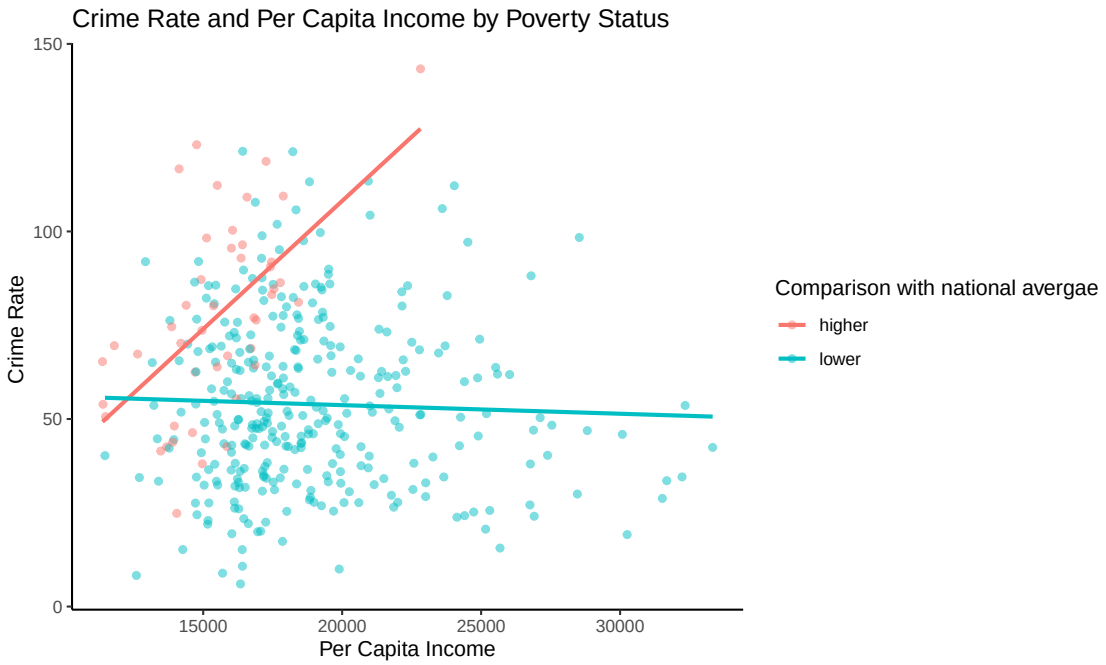


Figure 6: Interaction plot of Income Per Capita and Poverty

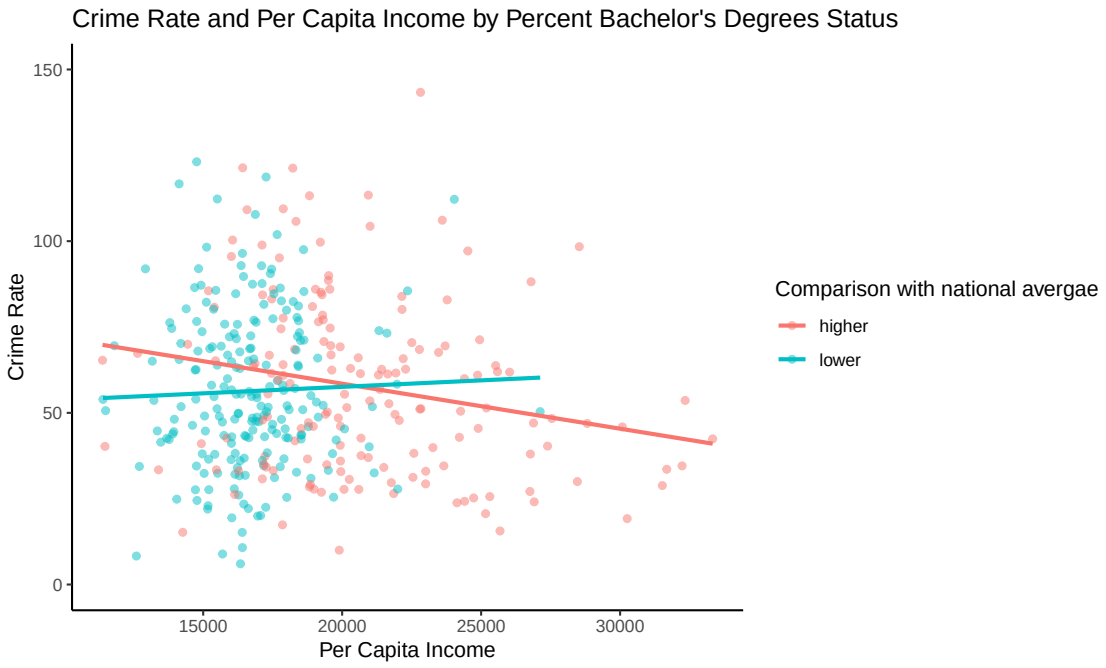


Figure 7: Interaction plot of Income Per Capita and Bachelor's Degree Status

Model Construction

By adding interative terms to the model built with selected variables, we resulted in the following three models:

Model 1 was built using the following terms: pop, pop18, bagrad, poverty, unemp, pcincome, pcincome*pop, region, beds_rate_1000, density.

Model 2 was built using the following terms: pop, pop18, bagrad, poverty, unemp, pcincome, pcincome*pop, region, beds_rate_1000, poverty*pcincome

Model 3 was built using the following terms: pop, pop18, bagrad, poverty, unemp, pcincome, pcincome*pop, region, beds_rate_1000, density, pcincome*bagrad

Model Diagnosis and Transformation

Using Variable Inflation Factor(VIF), we determined if each model had a high collinearity. The above 3 models all have passed the collinearity test.

We then drew diagnostic plots for each model to see how the residuals behaved. In all three models, residuals followed a normal distribution ($\mu = 0$) with no influential points.

In addition, we drew boxcox plots to see if each model needed transformation. The peak of all three boxcox plots fell between 0.5 and 1. As such, we tried $\sqrt{y}$ transformation for each model. Since the residuals were more unevenly distributed in all three transformed models, we kept the untransformed model. Detailed plots can be seen in main.Rmd.

Here is a summary table of for our three models:

Table 5: Models summary

	Model 1				Model 2				Model 3			
Coeffcient	*Estimates*	*std. Error*	*Conf. Int (95%)*	*P-Value*	*Estimates*	*std. Error*	*Conf. Int (95%)*	*P-Value*	*Estimates*	*std. Error*	*Conf. Int (95%)*	*P-Value*
Intercept	-83.18	15.49	-113.63 – -52.72	**<0.001**	-41.43	15.68	-72.27 – -10.59	**0.009**	-114.11	19.63	-152.72 – -75.51	**<0.001**
Total population	0.00	0.00	0.00 – 0.00	**<0.001**	0.00	0.00	0.00 – 0.00	**<0.001**	0.00	0.00	0.00 – 0.00	**<0.001**
Percent of population aged 18-34	1.25	0.32	0.62 – 1.88	**<0.001**	0.93	0.23	0.47 – 1.38	**<0.001**	1.05	0.33	0.40 – 1.69	**0.002**
Bachelors proportion	-0.37	0.25	-0.86 – 0.13	0.145					1.25	0.69	-0.10 – 2.60	0.069
Percent below poverty level	1.66	0.39	0.90 – 2.43	**<0.001**	-2.78	1.09	-4.93 – -0.62	**0.012**	1.91	0.40	1.12 – 2.69	**<0.001**
Percent unemployment	0.75	0.53	-0.30 – 1.79	0.160	0.94	0.48	-0.01 – 1.90	0.052	0.90	0.53	-0.14 – 1.94	0.091
Per Capita income	0.00	0.00	0.00 – 0.00	**<0.001**	0.00	0.00	-0.00 – 0.00	0.112	0.01	0.00	0.00 – 0.01	**<0.001**
Region-North Central	12.03	2.62	6.88 – 17.18	**<0.001**	10.79	2.57	5.73 – 15.85	**<0.001**	12.44	2.60	7.32 – 17.56	**<0.001**
Region-South	29.67	2.68	24.40 – 34.95	**<0.001**	27.69	2.64	22.50 – 32.89	**<0.001**	29.79	2.66	24.55 – 35.02	**<0.001**
Region-West	24.63	3.13	18.47 – 30.78	**<0.001**	21.32	3.03	15.35 – 27.28	**<0.001**	24.32	3.11	18.21 – 30.43	**<0.001**
Number of hospital beds per 1000 person	3.09	0.69	1.74 – 4.44	**<0.001**	1.99	0.72	0.56 – 3.41	**0.006**	2.74	0.69	1.37 – 4.11	**<0.001**
Density	0.00	0.00	-0.00 – 0.00	0.199					0.00	0.00	-0.00 – 0.00	0.134
Total Population*income per capita	-0.00	0.00	-0.00 – -0.00	**<0.001**	-0.00	0.00	-0.00 – -0.00	**<0.001**	-0.00	0.00	-0.00 – -0.00	**<0.001**
Poverty*income per capita					0.00	0.00	0.00 – 0.00	**<0.001**				
Bachelors proportion*income per capita									-0.00	0.00	-0.00 – -0.00	**0.012**
Observations	367				367				367			
R^2 / R^2 adjusted	0.563 / 0.548				0.580 / 0.567				0.571 / 0.555			

Cross Validation

We performed cross validation on each model and got the RMSE average. The results were shown in the table below:

Table 3: RMSE table for three models

model	RMSE	R Square
1	16.60	0.534
2	16.20	0.561
3	16.46	0.542

Model Assessment

We assessed the models we built in the testing set and evaluated them by R^2,$RMSE$ and $RMSPE$. The results were shown in the table below:

Table 4: Model assessment table

Model	R Square	RMSE	RMSPE
1	0.534	16.60	12.06
2	0.561	16.20	12.32
3	0.542	16.46	11.90

Conclusion and Discussion

According to table 5, evaluated in terms of accuracy in estimating the training set, model 2 has the best performance with its leading R square and the smallest RMSE, followed by model 3 and model 1. In terms of predicting values outside the training set, model 3 has the best performance with the smallest RMSPE on the testing set. Such an inconsistency of model performance may be explained by model overfitting of the training data. Above all, although model 2 have achieved a good estimation of the training set, we will choose model 3 as the final model, given its fair RMSE and R-square values, plus excellent testing set performance.

Overall, our project has several strengths. First, we did feature engineering before training the model by transforming variables using our domain knowledge to the ones more relevant to the predicted variable. For example, while area and population are not directly related to the crime rate, population density (population/area) can be more relevant. Second, we did analysis of correlation and collinearity to reduce potential bias. Third, we did interactive analysis and involved multiple interactive terms in our model, which to some extents represented the possible non-linear relations between the parameters and the predicted value. Finally, we separated a testing set from the dataset at the very beginning, on which we evaluated the performance of several models, addressing the potential predicting errors caused by model overfitting.

Meanwhile, the project also has its limitations. Essentially, we identified several wrong data points in the original dataset: for example, the population of Los Angles, an outstandingly large number, is not consistent with the number found on Wikipedia. Given that some of the data are mistaken,

further works can be done on correcting the dataset using external data sources. Furthermore, our regression model considered only linear and interactive terms, while some parameters could be better fitted using polynomials or exponentials. More sophisticated models can be used in the future to better estimate the data.

Reference and Documentation

[1] Committee on Law and Justice, et al. *Understanding Crime Trends: Workshop Report.* Edited by Arthur S. Goldberger and Richard Rosenfeld, National Academies Press, 2009. Accessed 11 December 2021.

[2] Rosenfeld, R., Vogel, M. & McCuddy, T. Crime and Inflation in U. S. Cities. *J Quant Criminol* 35, 195–210 (2019). https://doi.org/10.1007/s10940-018-9377-x

[3] U.S. Department of Commerce Economics and Statistics Administration, Bureau of the Census. Current Population Reports. Poverty in the United States: 1990. Series P-60, No.175

[4] Bureau of the Census. We asked... You told us. Census Questionnaire Content, 1990 CQC-13

[5] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

As Numerous new questions emerging during our discussion, our group explored materials below to solve them.

1. Question: Do we still need a test set when using k-fold cross-validation? Source: https://stats.stackexchange.com/questions/225949/do-we-need-a-test-set-when-using-k-fold-cross-validation https://datascience.stackexchange.com/questions/80310/is-a-test-set-necessary-after-cross-validation-on-training-set

2. Question: How to achieve build test set & predict Source: https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/ https://campus.datacamp.com/courses/machine-learning-with-caret-in-r/regression-models-fitting-them-and-evaluating-their-performance?ex=8

3. Question: How to evaluate continuous by continuous interactions Source: Continuous by Continuous Interactions, Joel S Steele http://web.pdx.edu/~joel8/resources/ConceptualPresentationResources/ContinuousByContinousInteractions_walkthrough_v2.pdf