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 implement several variable selection methods. Using the selected variables, we trained a linear regression model, elaborate 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. The third model has the best prediction on new data (in the testing set), whereas the second model, despite a better performance of the training data, may have the problem of overfitting. 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

Dataset discription

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.

Model building method

1. Data preprocessing:

Considering transformation of vairables in order to extract interpretable information form correlated predictors.

2. Exploratory analysis:

- Calculate the pairwise correlations between variables
- List all the correlations between the crime rate (our interest) and all other variables

3. Training/Testing set split: Randomly split the dataset into training and testing sets. 90% is training set while 10% is testing. This step aims to support model assessment and avoid overestimation.

4. Remove outliers and high leverage points

Use percentile to detect potential outliers and high leverage points. Due to the dataset size, we remove rows containing the smallest and largest 0.2% for each variables.

5. Model construction:

- Select variables using stepwise regression and criteria based procedure
- Build model using the variables we selected
- Plot interaction effects and add interaction terms
- Diagnose and transform the models

6. Cross validation

Cross validate on each model and get the model with the lowest RMSE.

7. Model assessment

In this section, we intend to assess the models we built and choose a final model used for future prediction. We compared 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 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 compute the number of crimes, physicians, and hospital beds per 1000 people. When considering crime rate, population density could be a key factor. Thus, we also create a new variable, `density`, which is population divided by area.

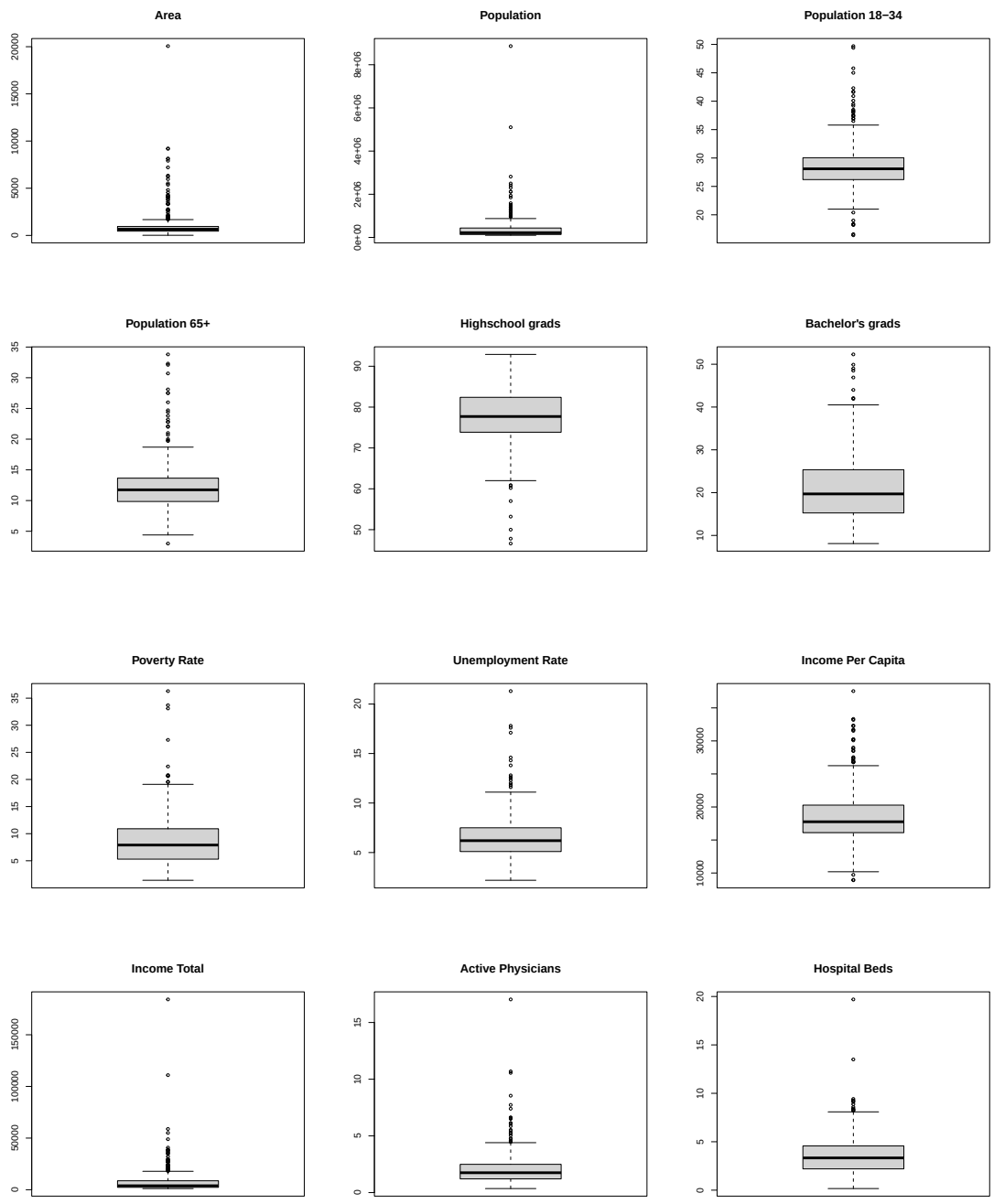


Figure 1: boxplot of continuous variables distribution

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:

We then take a closer look of each variables, calculate the pairwise correlations between variables,

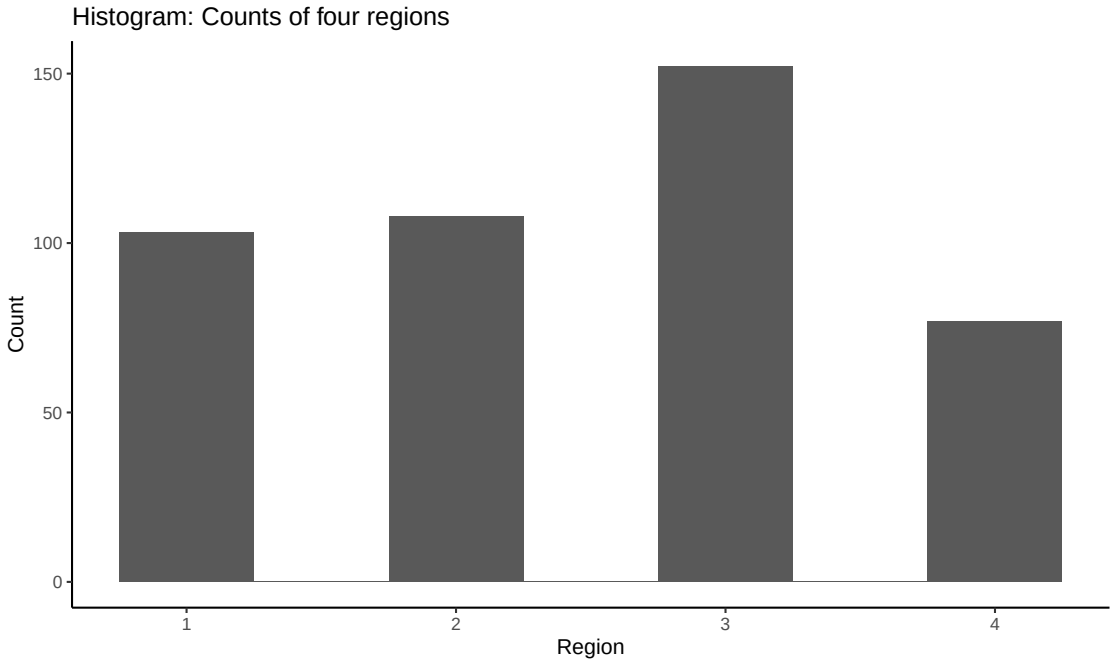


Figure 2: Histogram of catagorical variable:region distribution

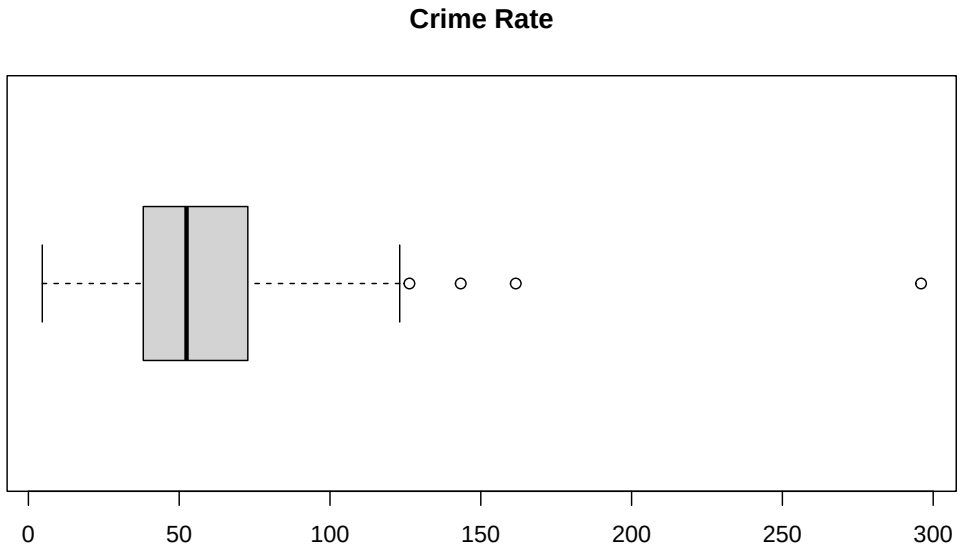


Figure 3: boxplot of dependent variable: crime rate

and list all the correlations between the crime rate (our interest) and all other variables. We used correlation analysis to prove that the derived variable density is a more meaningful variable compared to area and population, given that it has a stronger association with the crime rate.

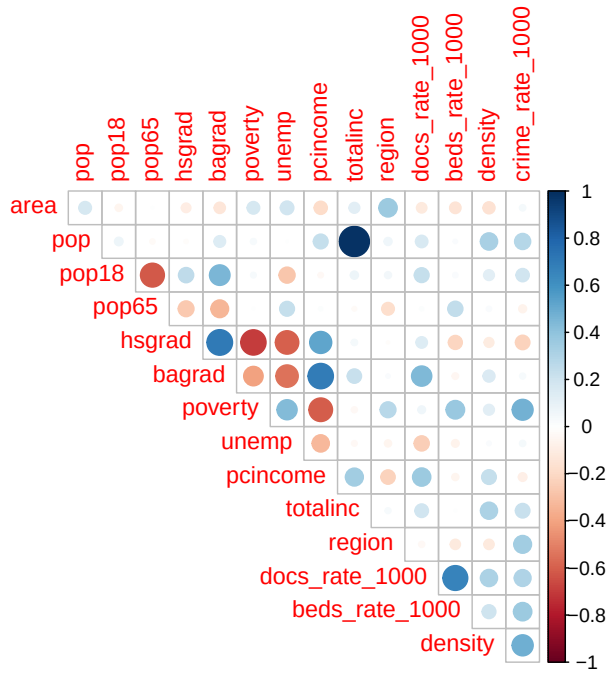
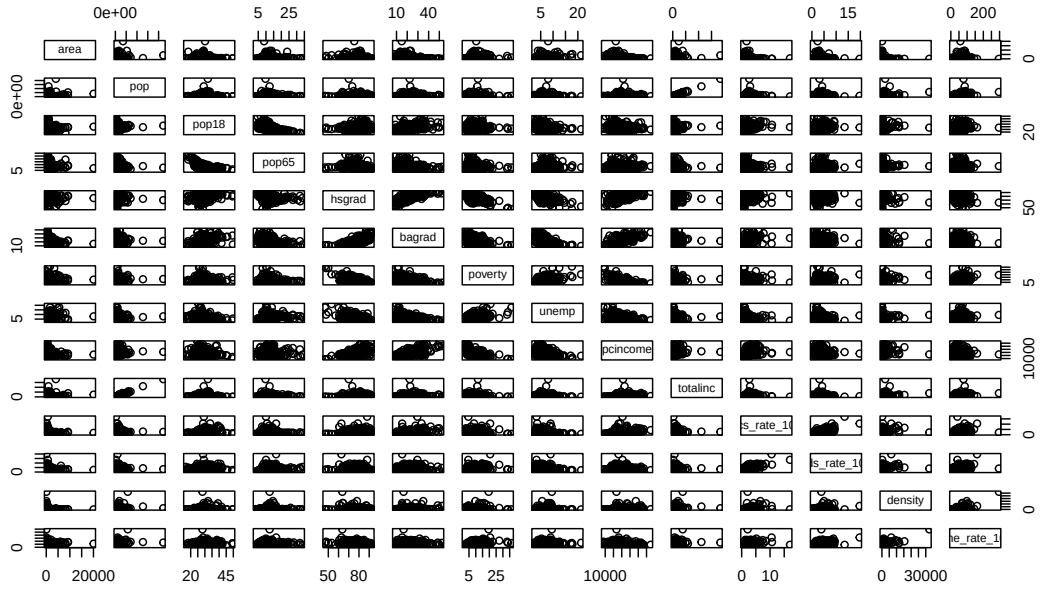


Figure 4: Correlation heatmap



After preliminary analysis of the data, we identify 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+
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 and made it a testing set(n = 44). The rest is training set (n = 396).

Data cleaning

Due to the dataset size, we removed rows containing the smallest and largest 0.2% for each variable. 19 data points were removed out of 396.

Model construction

1. Variables Selection

Based on stepwise procedure, we selected the following variables:

Table 2: Vairable selected from stepwise regression

backward	stepwise
pop	pop
pop18	pop18
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 find that `unemp` and `density` could also be selected.

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

2. Interaction effects

Interaction term 1: `poverty` + `income`

According to Census Bureau, the number of persons below the official government poverty level was 33.6 million in 1990, representing 13.5 percent of the Nation's population []. Thus, we can use

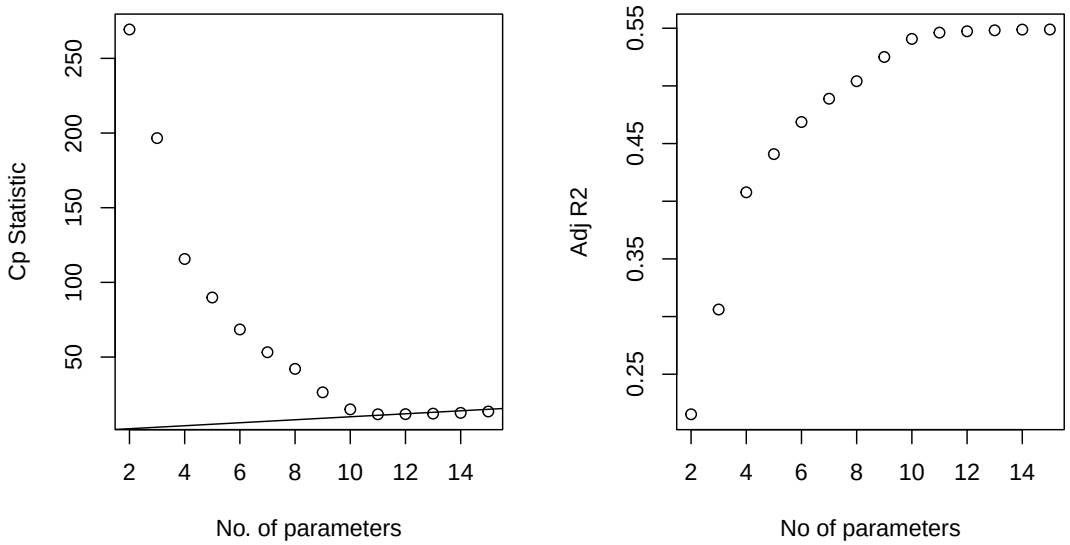


Figure 5: Subset selection for best parameter numbers

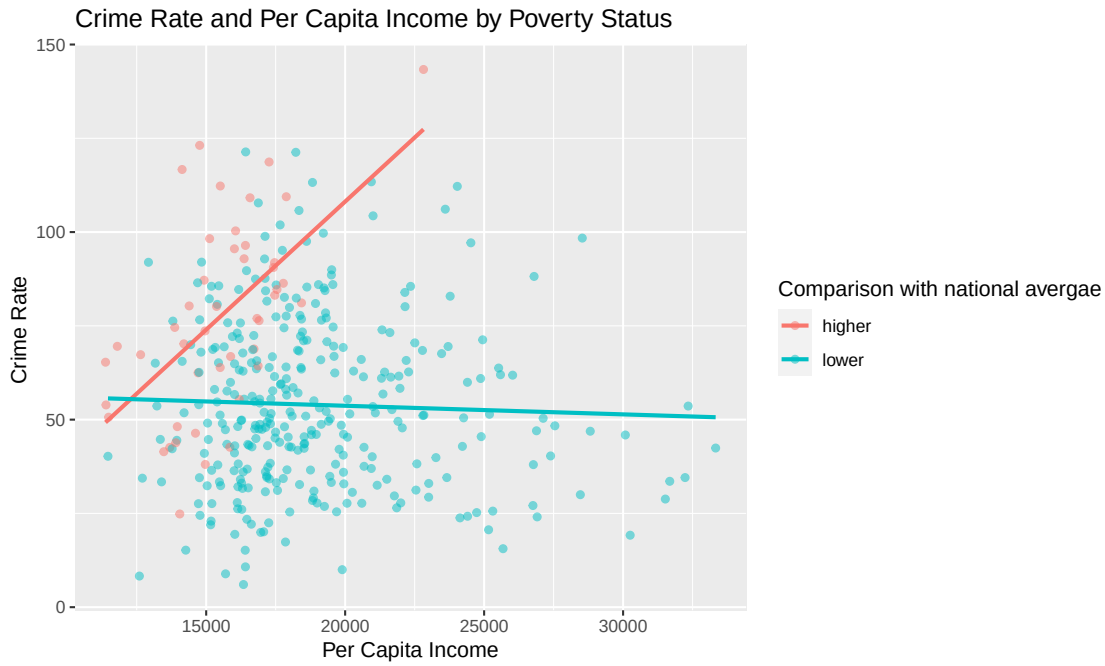


Figure 6: Interaction plot of Income Per Capita and Poverty

this criteria to divide `poverty` into two category: higher than national poverty rate and lower than national poverty rate.

Interaction term 2: `pcincome` + `bagrad`

According to Census Bureau, national percentage of persons 25 years old or older with bachelor's degrees is 20.8% [4]. Thus, we can use this criteria to divide `bagrad` into two category: higher than national `bagrad` and lower than national `bargrad`.

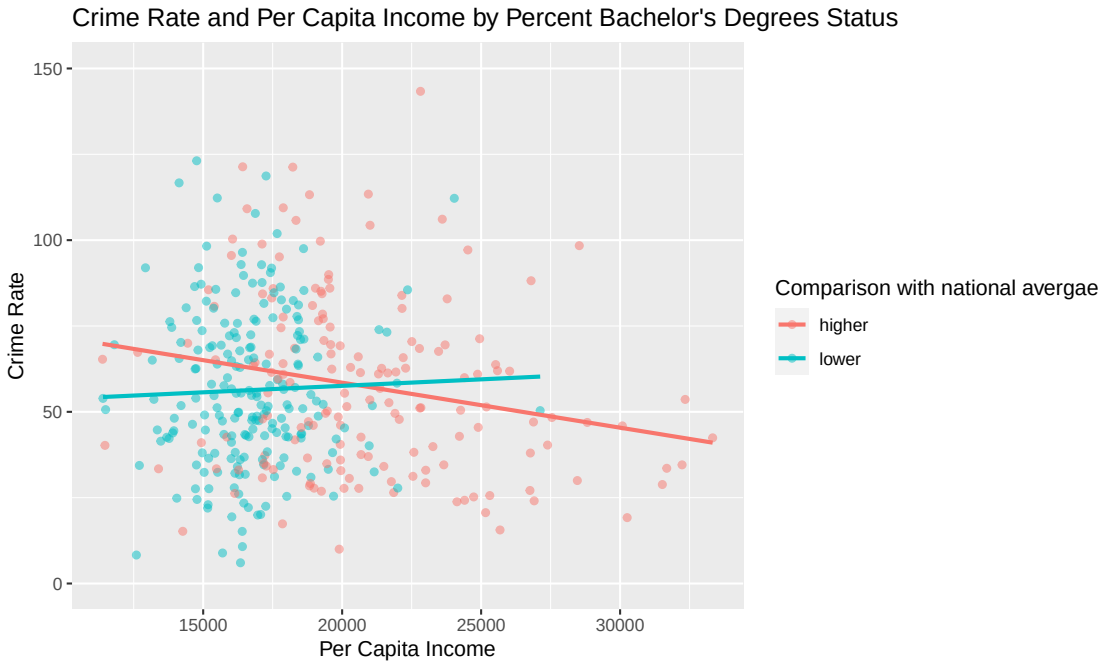


Figure 7: Interaction plot of Income Per Capita and Bachelor's Degree Status

3. Build model

Our model building criteria is removing variable with p-values > 0.2. We then came up with the following three models.

Our first model only using variables we selected:

$$crime\_rate\_1000 = pop + pop18 + bagrad + poverty + unemp + pcincome + pcincome * pop + regionbeds\_rate\_1000 + density$$

Our second model:

$$crime\_rate\_1000 = pop + pop18 + poverty + unemp + pcincome + pcincome * pop + region + beds\_rate\_1000 + poverty * pcincome$$

Our third model:

$$crime\_rate\_1000 = pop + pop18 + bagrad + poverty + unemp + pcincome + pcincome * pop + region + beds\_rate\_1000 + density + pcincome * bagrad$$

4. Model diagnosis and transformation

Use Variable Inflation Factor(VIF), we determine if each model has high collinearity. We choose models that only have low collinearity.

We then drew diagnose plots for each model to see how residuals behave. In all three models, residuals follows normal distribution with mean = 0, and there is no influential points.

In addition, we drew boxcox plots to see if each model need transformation. The peak of all three boxcox plots are close to 0.5~1. As such, we try $\sqrt{y}$ transformation for each model. Detailed plots can be seen in main.Rmd.

Compare to the diagnose plots of untransformed models, we found that the residuals are more unevenly distributed in all three models. Therefore, transformed models are worse. We selected the untransformed models.

Cross validation

We performed cross validate on each model and get the model with the lowest RMSE. The results are shown in the table below:

Table 3: RMSE table for three models

model	RMSE	R_sq
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 data and evaluated them by R^2,$RMSE$ and $RMSPE$. The results shows in the following table:

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 the above table 4, 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 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, which reduced the potential bias caused by confounders. 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 source. Furthermore, our regression model considered only linear and interactive terms, while some parameters could be better fitted using polynomials or exponents. 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