Deciphering the Determinants of Crime Rates

A Lasso Region Analysis of Socio-Economic Factors

Jiayu Ellie Su

Mentor: Prof. Nan Li
Department of Mathematics
CUNY-NYC College of Technology

Table of Contents

1. Introduction

Crime rates are influenced by a myriad of socio-economic factors, making them a complex phenomenon to understand and mitigate. In this project, we delve into the task of deciphering the determinants of crime rates using a Lasso Regression approach in R to identify and rank the top three factors that have the most impact on the criminal rate. By analyzing a comprehensive dataset encompassing various socio-economic variables, from the Communities and Crime database from the UC Irvine Machine Learning Repository **[1],** we aim to identify the key factors driving violent crime rates in communities.

This paper has 2 key chapters. Chapter 1, Methodology, delves into the core techniques used to complete this project. It includes the Lasso regression definition, what is variable selection, and the process to complete this project.

Chapter 2, titled Data Analysis and Conclusions, will show both the data preprocessing steps and the outcomes of the linear models, with a specific focus on identifying and ranking the top three factors that have the most impact on the criminal rate according to the best lasso model retained.

In summary, this project aims to intricate the relationship between socio-economic factors and criminal rate. By using lasso regression, the aim is to extract the most impactful predictors from the Communities and Crime dataset.

2. Methodology

Variable selection refers to the process of identifying and choosing the most relevant predictor variables and removing the unusable ones. Not all variables should be incorporated into the analysis due to their different capacities to provide informative insights. For example, it doesn't make sense to add zip codes to a linear model as they might be an individual ID, and also cannot be treated as numbers on which standardization can be performed. Variables with a high degree of missingness should also be removed, as they do not provide enough information to reflect their relevance to the outcome variable.
The steps that we took to identify and rank the top 3 factors that impact crime rates:

1) Download the Communities and Crime data from the UC Irvine Machine Learning Repository **[1]**.
2) Variable selection. This includes handling missing values, and outliers, and picking out variables that cannot be used (variable selection).
    a. Missing values: All variables with a missing rate higher than 0.2% were removed. For variables with a missing rate lower than that, the median was imputed to

substitute missing values - it was the case for the variable OtherPerCap only, with a single missing value.

b. Variables removal: The variables *state*, *county*, *community*,
c. *communityname*, and *fold* were removed from the dataset due to the inability to use a random training/testing split on those variables without affecting the results.
d. Data standardization: All numeric variables were standardized to the z-score scale to provide better interpretability to the model results.

3) Build and compare linear models with LASSO regularization for crime rates.
   a. Split Data into Train and Test Sets: The dataset is divided into training and testing sets to facilitate model evaluation. This ensures that the model's performance is assessed on unseen data.
   b. Fit Lasso Regression Models with Different Alpha Values: Lasso Regression models are fitted using the *glmnet* package in R, with varying alpha values. Alpha controls the degree of regularization applied, where higher values of alpha lead to more shrinkage of coefficients towards zero. Then, each model is trained on the training dataset.
   c. Model Evaluation: Cross-validation is performed to assess the performance of each model. We use k-fold cross-validation with k=10 to obtain robust estimates of model performance. Then, the Mean Squared Error (MSE) is calculated for each model. MSE measures the average squared difference between the actual and predicted values of the target variable (violent crime rates).
   d. Model Selection: The model with the lowest MSE is selected as the best-performing model. This model strikes the best balance between bias and variance, providing the most accurate predictions while avoiding overfitting. Additionally, the optimal regularization parameter (lambda) for the selected model is determined. Lambda corresponds to the penalty term applied to the coefficients in Lasso Regression, controlling the degree of regularization.
   e. Interpretation and Visualization: The results are interpreted to understand the predictive performance of the selected model and the importance of different predictors in explaining crime rates.

3. Data Analysis and Conclusions

3.1 Variable Selection

The dataset initially consisted of 128 variables, including *ViolentCrimesPerPop*, the dependent variable of this analysis.

To prepare the dataset for modeling, the following steps were undertaken:

1) Removal of Inappropriate Variables:

- Variables such as *state, county, community, communityname*, and *fold* were deemed unsuitable for model evaluation and hence were removed from the dataset.

2) Handling Missing Data:
   - Missing data were evaluated, and variables with a missingness rate exceeding 0.2% were eliminated (n = 22).
   - Only one feature, *OtherPerCap*, had a small rate of missingness, with a single missing value. Imputation using the median value was performed for this variable.
3) Data Standardization:
   - All numeric variables were standardized to z-scores to ensure uniform scales across variables. This standardization facilitates better interpretation of model coefficients and comparisons between predictors.

After variable selection and preprocessing, 100 variables remained for evaluation as potential predictors of crime rate (*ViolentCrimesPerPop*).

3.2 Lasso Regression Models

Two different approaches were conducted for different Lasso Regression models. The first model had a data split of 80% training and 20% testing, and the second consisted of a 70% training and 30% testing split. Below the results of each one will be described.

3.2.1 Model 1 - 80% training

The first step of a Lasso regression is to find the minimum lambda for the model, given the training data. A cross-validation was performed for this purpose with 10 folds. The resulting plot is depicted below. The analysis revealed 75 predictor variables with non-zero coefficients.

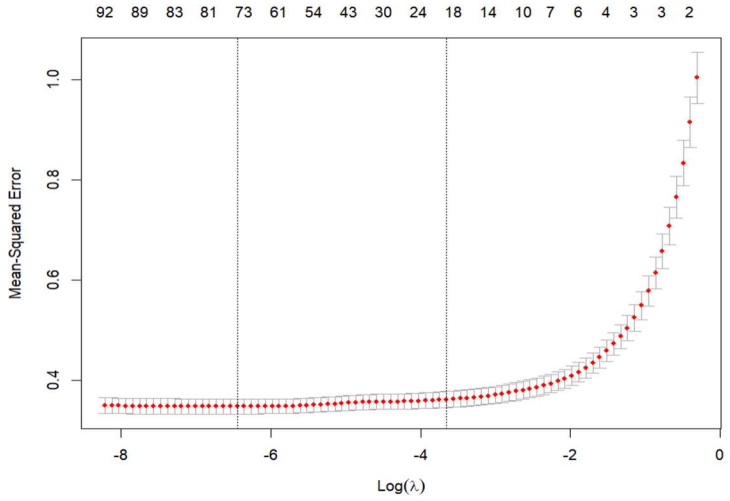

Figure 1. 10-fold cross-validation to find minimum lambda of Model 1.

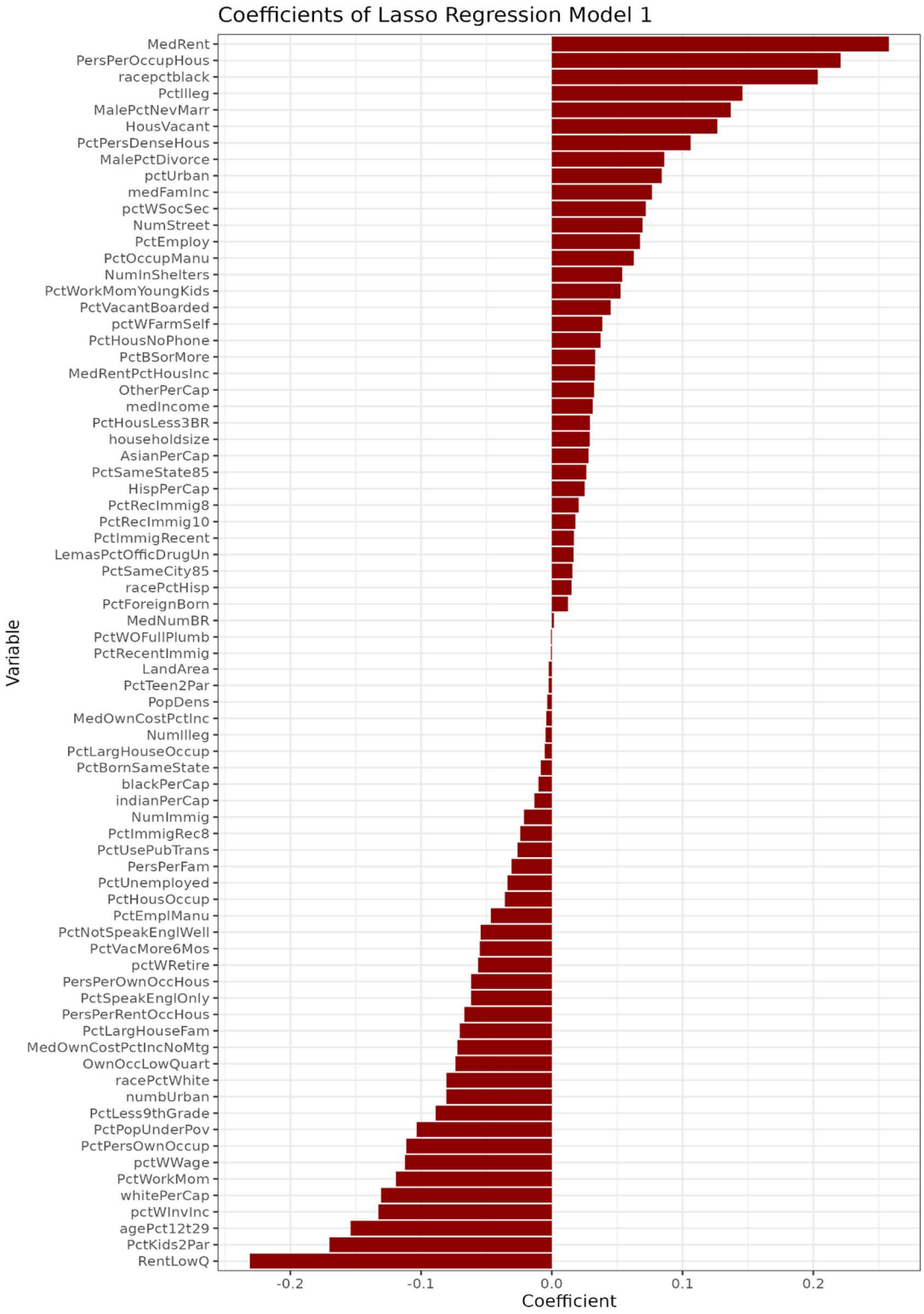

Figure 2. Non-zero variable coefficients for Lasso Regression Model 1.

The model has an MSE (Mean Squared Error) rate of 0.305, and an R^2 of 0.685, which means it explains 68.5% of the crime rate variance.

The table below shows the top 3 factors that impact crime rates.

variable	coefficient
MedRent	0.26
RentLowQ	-0.23
PersPerOccupHous	0.22

Table 1. Top 3 factors that impact crime rates in Model 1.

The signs of the coefficients show the direction of the relationship between these factors and crime rates. The variables with positive coefficients (MedRent and Racepctblack) mean that an increase in those factors is associated with increased crime rates. For variables with negative coefficients (RentLowQ), it means the opposite effect. So, variables with larger coefficients mean they contribute more significantly to the outcomes in the Lasso Regression model.

"MedRent" denotes median rent values, including utilities, and its inclusion among the top factors suggests that housing affordability might play a pivotal role in increasing crime rates. Higher housing prices may indicate areas with higher living costs. In such areas, people who face financial challenges may resort to crime out of desperation or lack of viable economic opportunities. Also, areas with higher rental costs might have more valuable items in their houses. This could attract burglary or theft. So, if MedRent is higher in the area, the crime rates are too high.

"RentLowQ" refers to a variable that represents the low-end range of rental prices in a given area. (The bottom 25% of rental prices.) Surprisingly, this shows a tendency of places with the highest bottom quartile rental prices to have the lowest crime rates, at first glance being the opposite of what the previous variable shows. This may make crime rates go down because lower rental prices in a community could show more affordable housing options. This may attract people who may have faced economic challenges in a higher rent area. Economic stability can correlate with lower crime rates because people who are more economically stable are less likely to engage in criminal activities. This shows if RentLowQ is lower in the location, so are crime rates.

"PersPerOccupHous" represents the average number of persons per household. This variable presents a positive coefficient, indicating that as the average number of people per household increases, the crime rate tends to increase too. It might relate to the fact that people with lower economic power share their homes with more members of the family. Therefore, a neighborhood where more people are living together in the same household might have increased crime rates due to economic challenges faced by these people.

It is important to notice that all coefficients have the directions they are showing considering all of the other factors, and not as isolated variables. So, the interpretation provided here is given considering the joined influence of all variables together.

3.2.2 Model 2 - 70% training

The cross-validation to find minimum lambda has also been performed with 10-folds after splitting the dataset into 70% training and 30% testing.

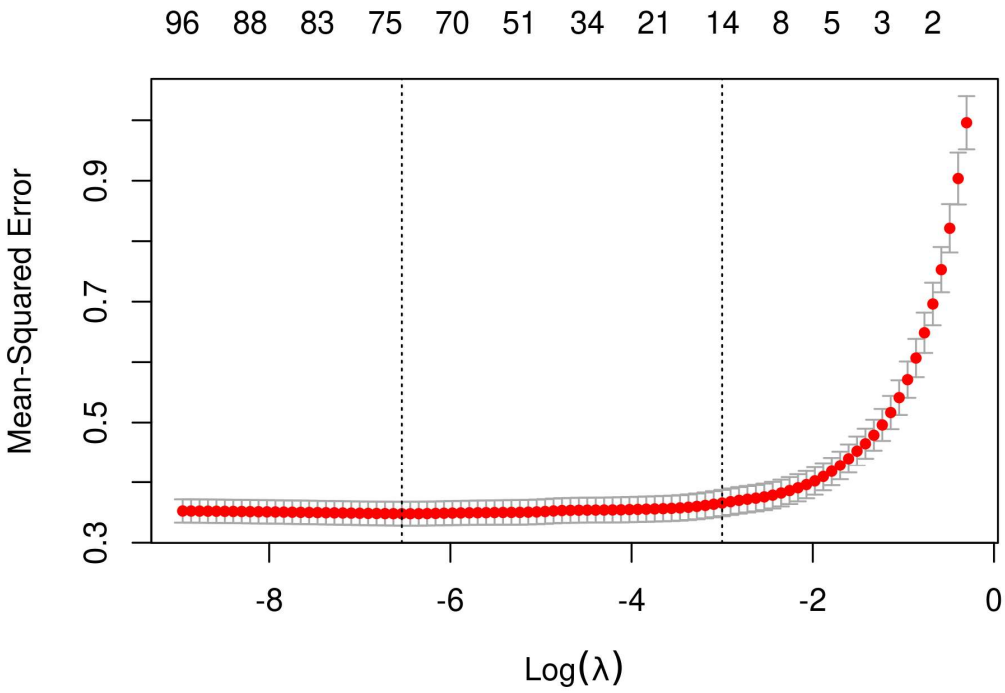

Figure 3. 10-fold cross-validation to find minimum lambda of model 2.

This analysis revealed 75 predictor variables with non-zero coefficients, as in model 1.

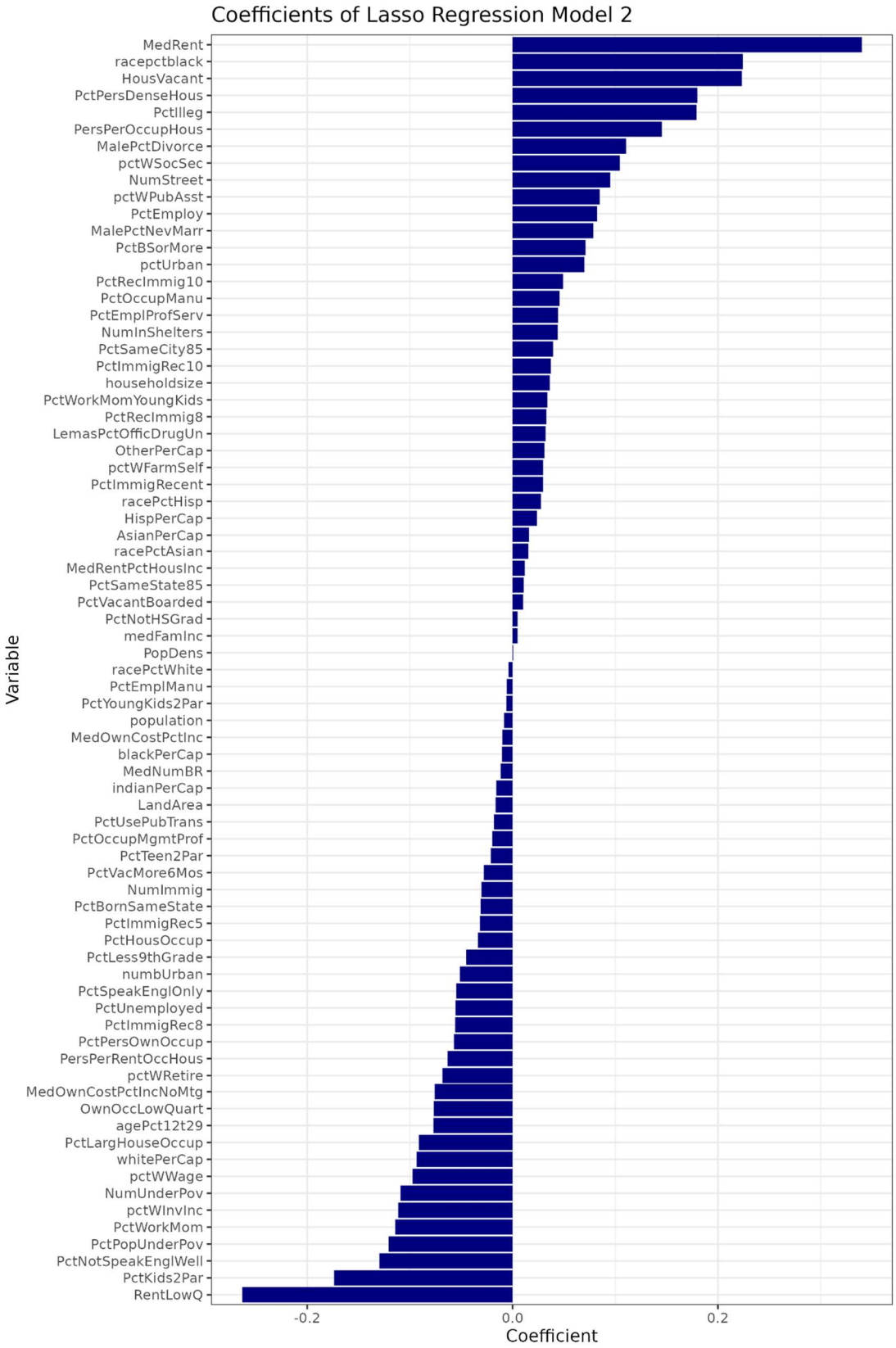

Figure 4. Non-zero variable coefficients for Lasso Regression Model 2.

The model has an MSE (Mean Squared Error) rate of 0.351, and an R^2 of 0.650, which means it explains 65% of the crime rate variance.

The table below shows the top three predictors for the crime rate of this model.

variable	coefficient
MedRent	0.34
RentLowQ	-0.26
racepctblack	0.22

Table 2. Top 3 factors that impact crime rates in model 2.

The two first coefficients are the same for model 1 and with the same direction. “Racepctblack”, the third most important coefficient, reflects the percentage of the population identifying as Black, which also is one of the top factors. This shows the significance of social characteristics in a community and how they might be connected to increasing crime rates. Communities with a higher percentage of black residents might also experience limited access to quality education, employment opportunities, and social services. These disparities can contribute to higher crime rates. This means in the location if Racepctblack is a higher percentage, so are crime rates.

As explained for model 1, for model 2 it is also important to notice that the results and interpretation are related to the influence of all variables together, in this multivariate Lasso linear regression model.

4. Conclusion

The MSE rate for model 1 is lower (0.305 vs. 0.35 for model 2), and the variance explained is bigger (R^2 0.685 for model 1 vs. 0.650 for model 2). Therefore, model 1 with a split of 80% for training and 20% for testing is the better choice to identify the most important variables related to crime rate in the analyzed dataset.

In conclusion, after creating the Lasso Regression model, our data analytics have found the top three factors that contribute to crime rates the most. They are MedRent, the median rent and utilities value; RentLowQ, the bottom quartile of rent values; and PersPerOccupHous, the average number of persons per household. The full model explains 68.5% of the variability in crime rates. These findings underscore the intricate interplay of social and economic dynamics in determining

crime rates in a specific location. Addressing underlying socio-economic disparities and promoting access to affordable housing may help mitigate crime rates and foster safer communities.

References

[1] UC Irvine Machine Learning Repository Communities and Crime data
https://archive.ics.uci.edu/dataset/183/communities+and+crime

[2] Kumar, Dinesh. "A complete understanding of lasso regression" Great team learning, May 30, 2023
https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters).