Understanding US Crime Data by Feature Analysis

Chen Liu
Department of Mathematics
cliudh@connect.ust.hk

Hangyu Lin
Department of Mathematics
hlinbh@connect.ust.hk

Abstract

In this report, we will analyze the US Crime Data(1) which contains crime rates and other data in 59 US cities during 1970-1992. There are 36 variables in total containing numbers of 7 types of crime events in 59 cities during 1970-1992, e.g., murder, rape, etc, and other information like population, and the number of policies, etc. Based on the Crime data, we want to find which features or factors affect the number of crime events most and how these features affect the crime. To analyze features, we will use several methods including PCA, J-S & MSE estimator, Lasso, and Tree-based model. For each method, we study how the features influence the number of 7 types of crime events. Though different methods have different results, there is some common observation that the economic level, unemployment rate, and population of different ages and black affect the crime data most.

1 Introduction and Description of the US Crime Data

First of all, we introduce the basic information for the US Crime dataset and try to give a primary analysis of the data. This data was assembled by Steven Levitt(1) for the paper ' Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime.'. As the paper title shows, the data was used for estimating the effect of police on crime. There are a total of 36 variables in the crime data which can be divided into several groups. The first group or the target group consists of the 7 types of crime events including murder, rape, robbery, assault, burglary, larceny, and auto. String or integer-based variables like city name, and city identifiers can be defined as discrete covariate variables while some continuous values like real income per capita in a state can be categorized as continuous covariate variables. Besides, there are some indicator variables that only have 0,1 values indicating some events or properties, for example, there is an indicator variable for indicating whether a city has 2/3/4-year mayoral terms. In our report, we aim to understand which features or factors affect crime and how they affect the crime like positive or negative influences. Specifically, for each type of crime, we want to find out which variables affect the change in the number of the kind of crime events. Before using some statistical methods or data reduction methods, we can first analyze the data based on some simple computation for the data.

Preprint. Under review.

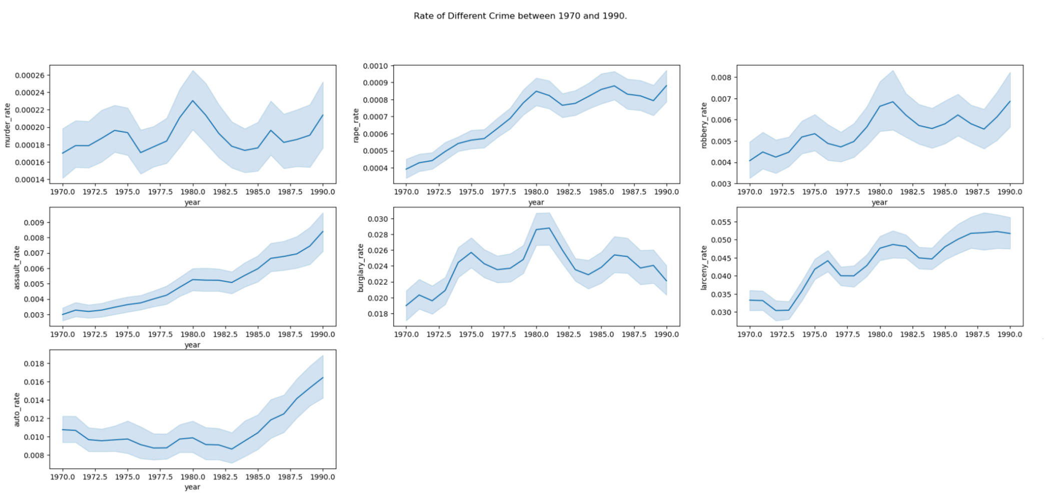

Figure 1: The mean and fluctuation of rates of 7 types of crime events in 59 cities during 1970-1992.

In Fig **??** and Fig 1, we plot the mean and fluctuation of the number and the rate (divided by the population of the city) with respect to 59 cities of each crime between 1970 and 1992. Some simple results shown in Fig **??** and Fig 1 show that there are rising trends for almost every type of crime as time goes on. Another observation from line plots is that the increase in the number of crime events is not only caused by the increase in population, but other factors affect the crime rate. Though the above results indicate that the rise of crime is not so closely related, different proportions of the population may affect crime in some way.

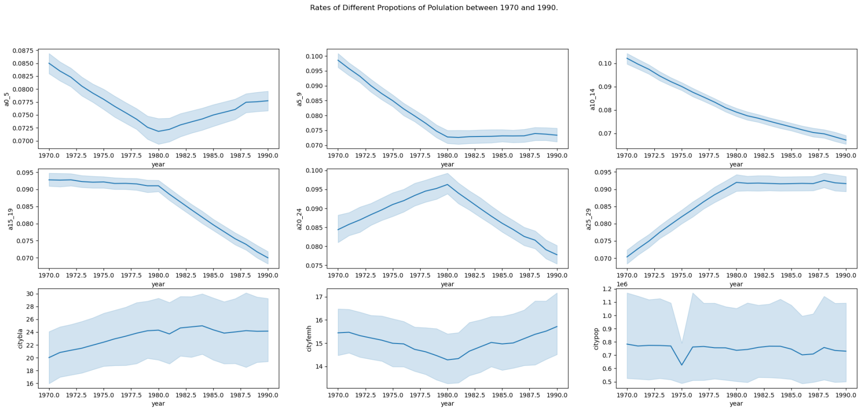

Figure 2: The rates or numbers for different parts of the population during 1970-1992.

As shown in Fig 2, there are different trends for different ages from 0-29 while none of them fits with the trend of the number of some kind of crime. In another way, the population of black seems to have a similar trend with the number of some crime events. Though it may be not the reason for the rise of crime, there may be some relationship between crime and the population of black. Since this is just a rough view and not a rigorous derivation, we can not conclude any statements until now. We can still do a similar analysis for other factors, but it can not be an efficient method to find out the features that affect crime. Further analysis will give more concrete and plausible results about this question.

2 PCA-based Feature Analysis

A possible way to analyze crime data can be PCA analysis, we employ the PCA on different combinations of variables, e.g., numbers of crime_types and indicator variables, numbers of crime_types and discrete covariate variables, etc. With the results of explained variance ratio in Fig 3, we can find for almost all combinations of variables, only one component makes the principal role which shows it is not a good way to directly use PCA method to do feature analysis.

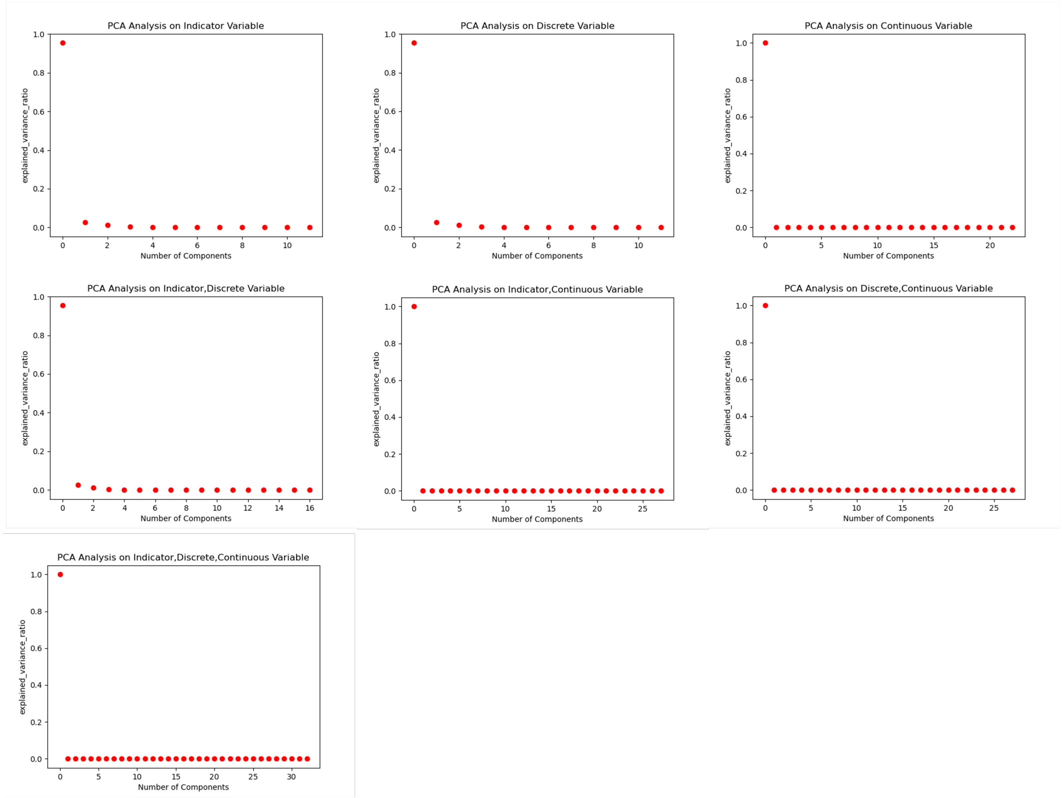

Figure 3: The PCA analysis that based on different groups of variables.

2.1 Horn's Parallel Analysis

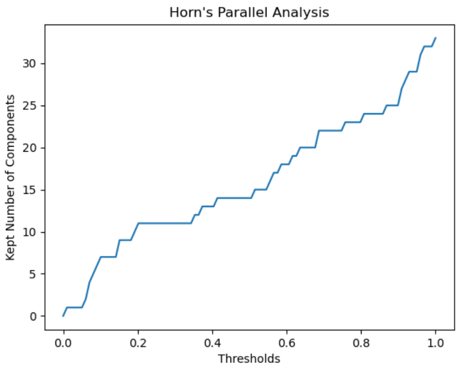

Figure 4: The Horn's Parallel analysis for all variables.

Furthermore, we also give the Horn's Parallel Analysis in Fig4, the results also indicate that PCA does not serve as a good method for Crime data. If we take the threshold as 0.05, there will be only one component left. Based on the above observations, we then try to use MSE or James-Stein estimators to find out which features make more important roles in determining the trend of crime.

3 James-Stein & MSE Estimator and Feature Analysis

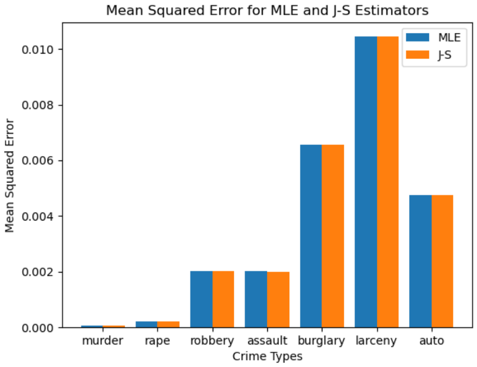

Figure 5: Mean Squared Error for MLE and James-Stein Estimators.

To deploy MSE or J-S estimators, we will take the rate of each crime type as the target and denoted by Y, all other variables rather than the number of crime events as covariate X. For each crime type, we fit (X, Y) by MLE and James-Stein estimators under the loss of mean squared error (MSE). The formulas for these two estimators are,

$$\hat{\beta}^{MLE} = (X^T X)^{-1} X^T Y \tag{1}$$

$$\hat{\beta}^{JS} = \hat{\beta}^{MLE}\left(1 - \frac{(p-2)\hat{\sigma}^2}{(\hat{\beta}^{MLE})^T X^T X \hat{\beta}^{MLE}}\right) \tag{2}$$

We use 10-fold cross-validation to evaluate the performance of these two kinds of estimators. As shown in Fig 5, we can find there is almost no difference in MSE values of MLE and J-S estimators (actually about 0.00001 fluctuation). Since the $n = 1000$ is much larger than $p = 26$ in this problem, the derivation of the Jamse-Stein theorem indicates that no significant improvement of J-S estimator is reasonable.

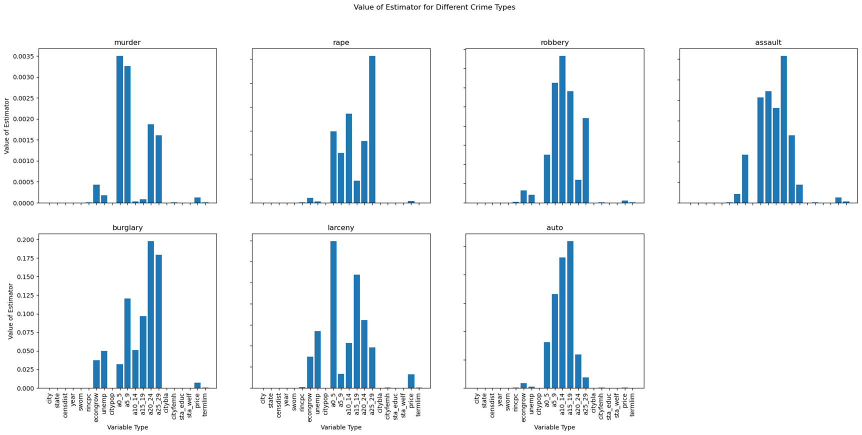

Figure 6: The Values of Estimator for Different Crime Types.

The find which factors affect the number of crime events most, we give the value of estimators to represent the importance of the corresponding variable/factor. From the results in Fig 6, we can find populations of different ages and black often take important roles in predicting crime data. Other factors like economics level and employ rate make a relatively large effect on some crime types, e.g., robbery, and assault.

4 Lasso

To further analyze which factors are dominant for the crime rate. We try to use the feature selection method of linear models. In specific, we use Lasso (3) to select useful features. Suppose that we have design matrix $X \in R^{n \times m}$, where n and m denote the number of samples and features. And we want to learn the parameter $\beta in R^m$. The target is $Y \in R^n$, consider a regression problem with square loss, the problem can be formulated as,

$$min_{\beta}\frac{1}{2}\|X\beta - Y\|_2^2 \tag{3}$$

For Lasso, it considers feature selection, the difference of formulation lies in the penalty term,

$$min_{\beta}\frac{1}{2}\|X\beta - Y\|_2^2 + \lambda\|\beta\|_1 \tag{4}$$

Here, λ is the hyperparameter for us to tune. Intuitively, if one feature stays selected for a large λ, it could be more likely to be the important feature. So we try to use Lasso to fit the relationship between the features and the cirme rate, and we draw the regularization path the Lasso to figure out the importance of each feature. Firstly, we consider the assult crime type. As shown in Figure. 7, sta_educ, citybla, sta_welf turn out to be important features. The real meaning for sta_educ is the combined state and local spending per capital on education. citybla stands for the percent of the city population that is black. sta_welf means the combined state and local spending per capital on public welfare. The spending in public welfare and education can affect the living standards of common people, so it is reasonable for them to be important features. For the percent of black people, it may be caused by some historical reasons.

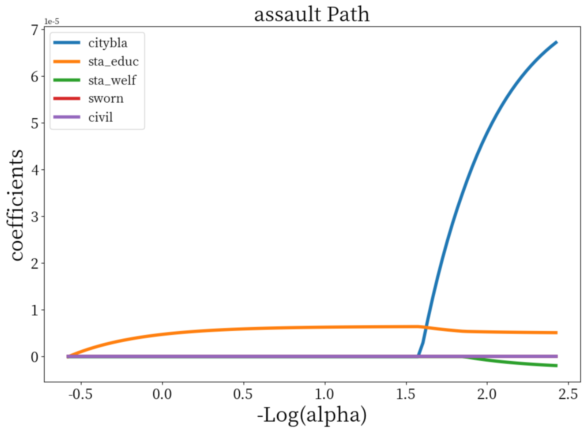

Figure 7: The lasso path for assault crime. The x axis is the $-log(\lambda)$ and y axis is the magnitude of the paramters.

The lasso paths for the remaining 6 crime types are shown in Figure.8. By observation, we can find that, for feature selection with Lasso, the important feature for these crime types are very similar, the combined state and local spending per capita on education and the percentage of the city population that is black are always important features.

5 Tree-based Model

To further explore the feature importance. We apply one commonly used non-linear method, tree-based model (2) to fit the relationship between the features and the target. For tree based method,

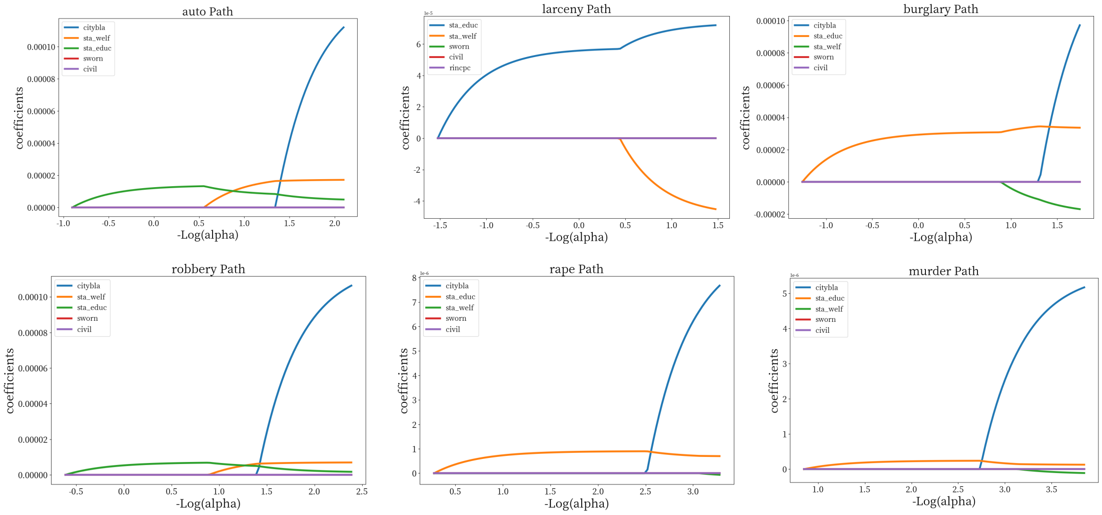

Figure 8: The lasso path for the remaining crime types. The x axis is the $-log(\lambda)$ and y axis is the magnitude of the paramters.

we can get the times of split for each features, which can be considered as the metric for feature importance. Here we visualize such kind of importance for further study. Firstly, we consider the assult crime type. The figure is shown in Figure. 9 In the barplot, we can find that civil, a15-19, citybla, cityfemh are relatively important features. Here civil stands for the number of civilian police employees by city. It can partially illustrate the police strength of the city. For cityfemh, we think that the possible explanation is that female households are more likely to be viewed as targets. a15-19 means the percent of population within 15-19, kids within such range are more prone to be targets of assult or commit a crime.

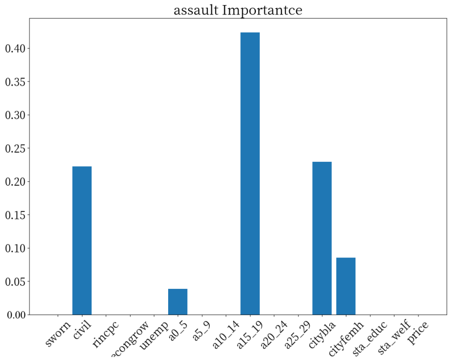

Figure 9: This figure shows the feature importance for the assault crime type.

The feature importance for the remaining 6 crime types are shown in Figure.10. By observation, we find that the percentage of black people is consistently an important factor. In addition, we find that sworn and civil are important features of tree model. The real meaning for them is the number of sworn police officers employed by the city and the number of civilian police employees by the city. The number of police employees in the city does affect the crime rate, so it is very reasonable.

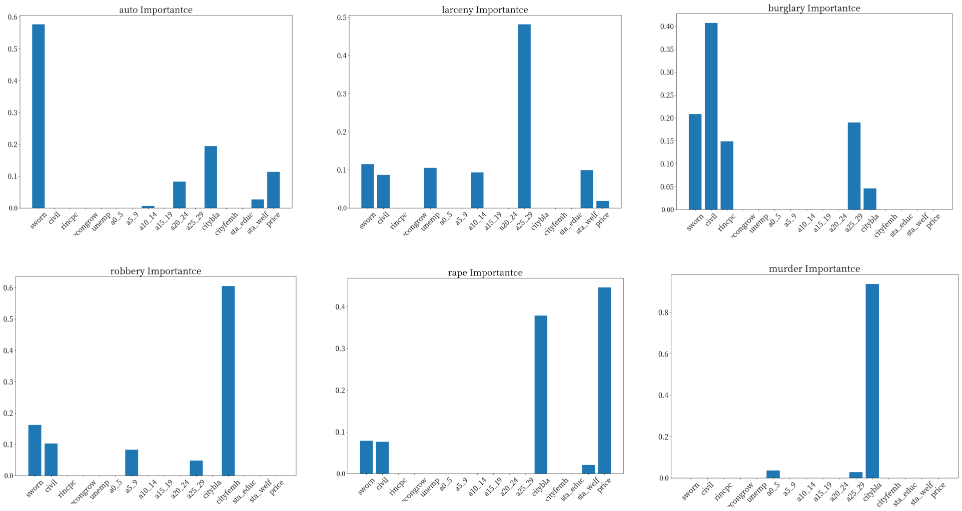

Figure 10: This figure shows the feature importance for the remaining 6 crime type.

6 Conclusion

In summary, we analyze the US Crime Data and attempt to find core factors that can predict the crime rate. First, after employing the PCA method, we find PCA is not so appropriate for this problem and the data. The results from MLE/James-Stein estimators show that economic level, unemployment rate, and the population of different ages and black affect crime during 1970-1992 most. In addition, we apply both Lasso and Tree models to select important features. For Lasso, the percentage of black people and capital spent on eductaion and welfare are selected as core factors. On the other hand, the number of sworn and civil police officers employed by the city and the percentage of black people is important for tree models. Here the difference may be caused by the nonlinearity of the tree model.

References

[1] LEVITT, S. D. Using electoral cycles in police hiring to estimate the effect of policeon crime, 1995.

[2] LI, B., FRIEDMAN, J., OLSHEN, R., AND STONE, C. Classification and regression trees (cart). *Biometrics 40*, 3 (1984), 358–361.

[3] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.