STAT 452 – Statistical Modeling I

Case Study 2

Paolo Furlanetto Ferrari

05/02/2023

Summary

I. Introduction

Combining geographical and social information with criminal data is an important component in predicting, diagnosing, and explaining criminal activity in communities, cities or even states. Thus, the statistical analysis of combined geographical-criminal datasets is a valuable tool for designing social policies. In this Case Study, we analyzed the geographical data of 1960 for 47 different states in the USA to try to predict their crime rates (number of offenses per 100,000 inhabitants) using Multiple Linear Regression. To achieve this goal, we have fitted and compared the performance of 4 different models using different variable selection criteria: with AIC criterion, with BIC criterion, PCA regression and Lasso regression.

II. Data Overview

The data consisted of 47 sets of 16 different variables for 47 different states in the US for the year of 1960. One such variable, the Crime Rate, is the variable we tried to predict using the other 15. To get an initial picture of how the data looks like, we plotted in Figure 1 the scatter plots of the Crime Rate versus each of the other variables. First, we can see that there is a lot of variation in most of the scatterplot and that it is difficult to see a trend in most cases. In most cases, the correlations with Crime Rate seem to be positive, although very weak. However, it is in fact possible to see a good correlation with some of the variables, notably the Police Expenditures in 1960 and 1959, the Unemployment Rates (classified between age group in years) and the Wealth (representing the median value of transferable assets or family income). While these conclusions are mostly visual, they offer a hint to the number of variables that should be included in the model selection algorithms tested next: roughly, we expect that only half of the variables or less will be relevant.

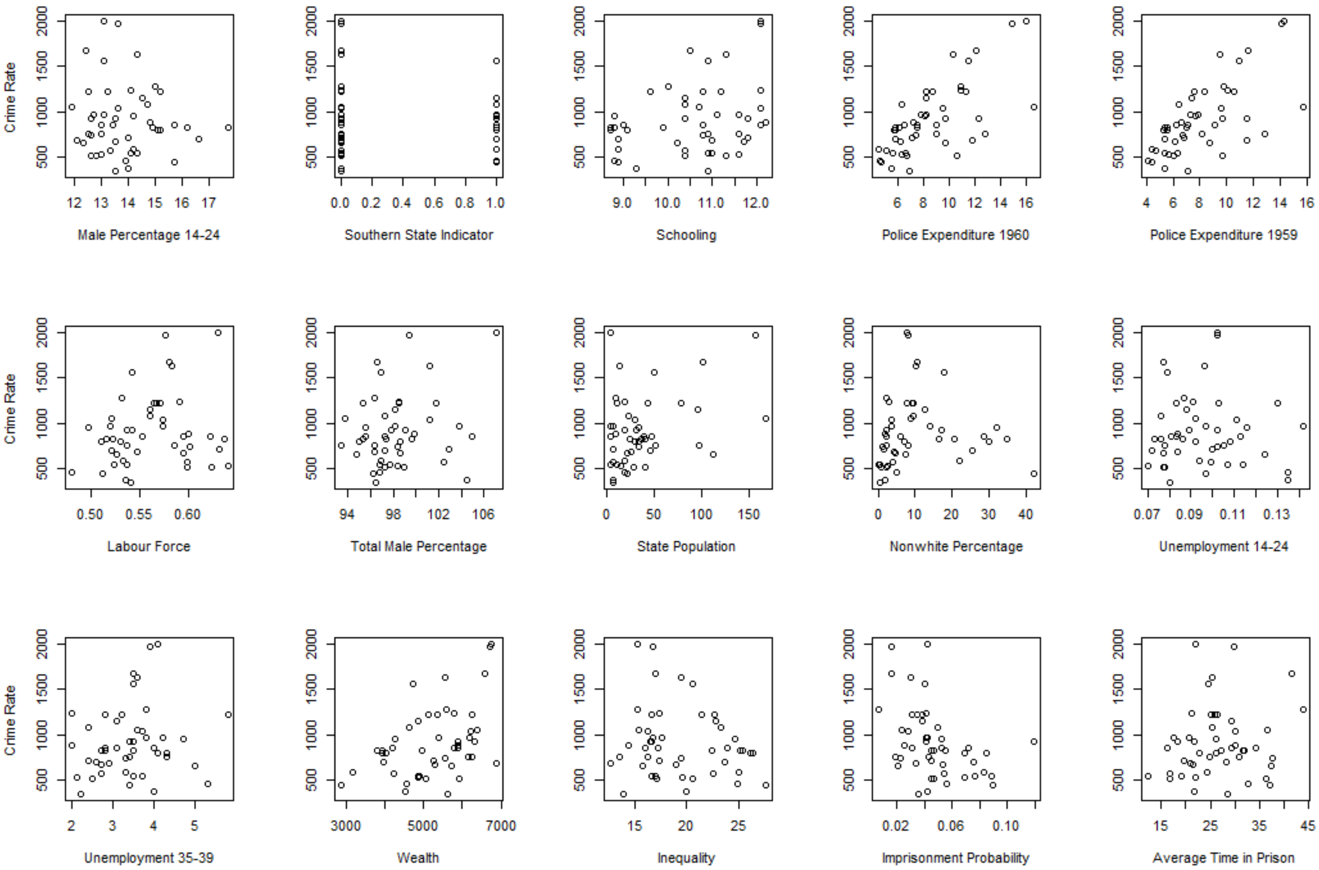

Figure 1: Scatterplots of Crime Rate (y-axis) vs. other scores in the dataset.

III. Data Preparation

To fit and evaluate the different models used in this Case Study, we have split the data roughly as 75% for training and 25% for testing. Since the total number of entries is 47, that amounts to 36 samples for training and 11 samples for testing.

The data preparation stage begins with loading the dataset into a dataframe in R and transforming it to numeric format. Then, we randomly assign 11 entries to a testing dataframe and 36 to training dataframes. We use the command “set.seed(1)” before splitting the data for reproducibility.

After preparing the data, we fit 4 different models using the train data according to different selection criteria: AIC, BIC, PCA and Lasso. After the 4 models are found, we compare their performance on the test data.

IV. Model Fitting

AIC and BIC:

The first 2 models are found using the AIC and BIC criteria for variable selection from the full Multiple Linear Regression model. For that, we use the Leaps and Bounds algorithm to detect which of the predictors should be included.

We find, for AIC, that 9 predictors should be included, namely: (1) Percentage of Males aged 14- 24, (2) Mean Years of Schooling for people aged 25 or over, (3) Per capita Police Expenditure in 1960, (4) Per capita Police Expenditure in 1959, (5) Number of Males per 100 Females, (6) Unemployment Rate for Males aged 14-24, (7) Unemployment Rate for Males aged 35-39, (8) Income Inequality (percentage of families earning below half the median income) and (9) Probability of Imprisonment (ratio of commitments to number of offenses).

For BIC, since it penalizes the total number of predictors more than AIC, we find that the best model is the one with only 3 predictors: (1) Police Expenditure in 1960, (2) Number of Males per 100 Females and (3) Income Inequality.

PCA:

PCA, or PCR, refers to transforming the predictors from their original format to their principal components, followed by a Multilinear Regression of the response variable with respect to these principal components. The idea is to try to explain the variation of the response from combined variations of multiple predictor variables, thus avoid collinearity issues.

In order to select how many principal components should be considered for the model, we have performed a cross-validation (CV) of the train data and have chosen the number of components with minimum mean squared prediction error across all folds. The number of folds for CV was 6, meaning that the train data (with 36 entries) was sliced in 6 parts, thus with 6 entries each. At each iteration, the model is trained in 5 of the 6 parts and the prediction error is computed on the remaining one.

The best PCA model (the one with less mean squared prediction error) contains 8 components, thus about half of the number of predictors.

Lasso:

Lasso is a regularization technique which penalizes Linear Regression Models with large coefficients. In order to fit a Lasso model, the penalization parameter λ needs to be specified. In general, the value of λ is obtained through cross-validation.

In R, we use the “glmnet” package to fit a Lasso regression while also performing a cross-validation. Just like for PCA, we use a 6-fold CV procedure to find the optimal λ. Figure 2 shows the Mean-Squared-Error (MSE) of the CV procedure versus the tuning parameter λ. As we can see, the algorithm predicts an optimal log(λ) of about 3, corresponding to λ = 31.92.

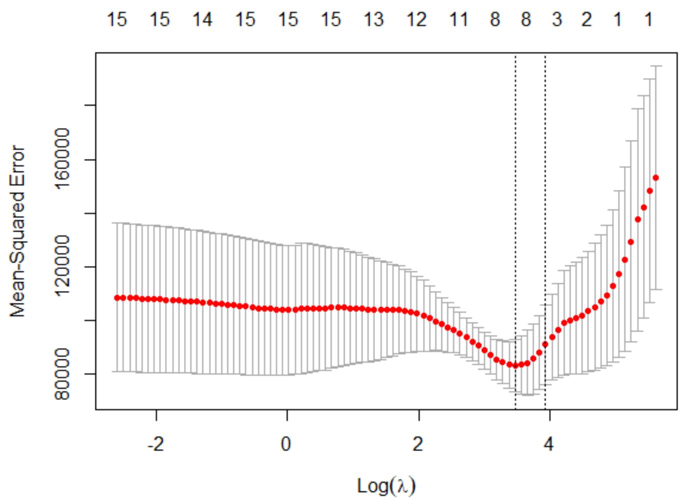

Figure 2: MSE of Lasso Regression using 6-fold CV.

V. Model Comparison

Finally, to select the best model out of the 4, we computed the MSE for each one using the test data. Table 1 shows the results. Based on these, we conclude that the model based on the AIC critetion has considerably lower error than the others and thus is the model of choice of our analysis.

Table 1: Mean-Squared-Error of prediction of the test set for each model.

	AIC	BIC	PCA	Lasso
RMSE	33804.57	71578.96	67222.03	83333.46

To summarize, in simple terms, our best model is the one based on AIC criterion. It can be written as

$$Y = \beta_0 + \sum_{i=1}^{9} \beta_i X_i + \epsilon$$

Where Y is the Crime Rate, and the variables X_i, for i = 1 to 9 are the variables representing:

(1) Percentage of Males aged 14- 24,

(2) Mean Years of Schooling,

(3) Police Expenditure in 1960,

(4) Police Expenditure in 1959,

(5) Number of Males per 100 Females,

(6) Unemployment Rate for Males aged 14-24,

(7) Unemployment Rate for Males aged 35-39,

(8) Income Inequality,

(9) Probability of Imprisonment.

VI. Conclusion

In this study, we combined criminal and geographical data of 47 states in the US of the year 1960. We concluded that the Crime Rates (number of offenses per 100,000 inhabitants) are mostly related to 9 different factors. In basic terms, these factors reflect the percentage of young males in the population, the mean years of schooling of the population, the ratio between males and females, the expenditures with police, the unemployment rate, the inequality of income and the ratio of prison commitments to offenses.

References

[1] Stat 452 Lecture Notes (2023).