Executive Summary

Case Study 2- Crime Dataset

Submitted By - *Debapratim Ghosh (dg19@illinois.edu) and Shantanu Solanki (solanki7@illinois.edu)*

Problem Description

Crime is one of the most burning problem in the society which need to be addressed. But the government (Federal or State) should not apply a one size fits all policy while tackling the burgeoning crime rates. Punishment regimes of the state along with citizen demographics play a major role in the extent of crimes in the state. Another important thing to understand is the interaction between offense and defense ; Crime and Collective Law enforcement. Finally, crimes can be influenced by the financial state of the citizen. The underlying hypothesis is that higher the inequality in income, higher the crime rates in the state.

Therefore, a group of criminologists have collated multiple variables associated with crime rates in 47 states. Our task is to understand which variables are important in accurately predicting crime rates and pick the **best** model with those variables.

Data Dictionary

The study conducted by the criminologists from University of Chicago utilised a dataset with the following columns :

`M` - percentage of males aged 14–24 in total state population
`So` - indicator variable for a southern state
`Ed` - mean years of schooling of the population aged 25 years or over
`Po1` - per capita expenditure on police protection in 1960
`Po2` - per capita expenditure on police protection in 1959
`LF` - labour force participation rate of civilian urban males in the age group 14-24
`M.F` - number of males per 100 females

`Pop` - state population in 1960 in hundred thousand
`NW` - percentage of nonwhites in the population
`U1` - unemployment rate of urban males 14–24
`U2` - unemployment rate of urban males 35–39
`wealth` - median value of transferable assets or family income
`Ineq` - income inequality: percentage of families earning below half the median income
`Prob` - probability of imprisonment: ratio of number of commitments to number of offenses
`Time` - average time in months served by offenders in state prisons before their first release
`Crime` - crime rate: number of offenses per 100,000 population in 1960 (**Response**)

As we can observe, the dataset includes variables related to the demographics of people, the financial condition of the citizens of the state ,the law enforcement expenditures and the punishment duration in the state.

Objective and Methods

The objective of this study is to identify the optimal set of predictors (or variables) which can help in predicting the crime rates in states with the highest accuracy. We would be using linear models to make our predictions on crime rates.

Along the way, we will explore different methods for isolating the set of predictors that gives the least prediction error.

Firstly, we will perform exploratory analysis on our dataset to check for the presence of unusual observations and linearly dependent variables. The presence of unusual observations would lead to a significant change in the regression structure of the model we are trying to build and hence they should be removed. Similarly, if there are linearly dependent variables, then the variance of our coefficients would be inflated and our estimation of model parameters would be inaccurate. So a few of the linearly dependent variables should dropped from our model.

Secondly, since we are primarily using linear models in predicting the crime rate, we should check if all the model assumptions are satisfied. These model assumptions include : Homoscedasticity (Or the errors have constant variance), Normality of errors and Linearity between the predictors and the response. If the first two model assumptions are not met, then we would have to transform our response using box-cox transformation or if the linearity assumption is not met, then we need to replace one (or more) of our predictors with a polynomial or any other transformation.

Finally, we would need to check the set of predictors which gives us the best prediction accuracy (or the lowest prediction error). For that we need to randomly divide our dataset into a training set where we "train" our model and testing set, where we "test" our model against unknown observations (or observations not previously exposed to the model).

We will explore multiple methods ranging from linear models , shrinking the coefficients using dimensionality reduction techniques or penalised regression or choosing the best model using algorithmic and criterion based variable selection methods.

Summary of Findings

1. From our exploratory data analysis, we didn't find any unusual observations in the dataset that might affect the regression structure of our model. We did find a few columns that are linearly dependent on each other **(Appendix A : Fig 1)** confirming the presence of **multicollinearity.** We removed `Po2` (per capita expenditure on police protection in 1959) and `Wealth` (median value of transferable assets or family income) from our dataset and re-fitted our model . This is done as the predictors removed don't add any extra information to the model. For example, `Po2` is highly correlated to `Po1` and therefore using `Po1` in the model is sufficient. Similarly, `Wealth` is negatively correlated to `Ineq` and therefore keeping `Ineq` is sufficient. Both `Po1` and `Ineq` turn out to be significant predictors in the linear model.
2. We found that none of the assumptions of the linear model were violated and the assumptions of homoscedasticity, normality and linearity were met.

3. Finally, to identify the model with the best prediction accuracy (Or **Test RMSE**) we explored multiple techniques and models. A summary of the prediction error obtained is given in the table below. A comparison of the prediction errors is given in **(Appendix A : Fig 2)**

	Model Used	Training RMSE	Testing RMSE
	Linear Model	146.035391	290.5508701
Searching Algorithm Based Selection	Forward Selection	171.9663207	178.9096836
	Backward Elimination	171.9663207	178.9096836
	Both Ways Selection	171.9663207	178.9096836
Criterion Based Selection	Adjusted R-Squared	155.1944661	324.0002465
	AIC	155.1944661	324.0002465
	BIC	162.7857041	320.9387219
	Mallow's Cp	155.1944661	324.0002465
Shrinkage Methods	Ridge	154.1485858	290.6275422
	Lasso	148.0356028	291.8799488
	PCR	152.6472062	309.8834303

Conclusion

Based on the prediction errors, we are choosing the model selected using Forward/ Backward or Bothways selection. The selected model has the following features

`So` -indicator variable for a southern state

`Ed` -mean years of schooling of the population aged 25 years or over

`Po1` -per capita expenditure on police protection in 1960

`M.F.` - number of males per 100 females

`Ineq` -income inequality: percentage of families earning below half the median income

`Time` -average time in months served by offenders in state prisons before their first release

The selected model explains 84% of the variation in the crime rate in the dataset. Additionally , predictors like education, law enforcement expenditure, % of males, income inequality and Time served by prisoners are all significant in predicting the crime rates.

The average prediction error of our model is 178.9 (Root mean squared error) which is the lowest among all the models considered.

The regression summary given in **(Appendix A : Fig 3).**

Appendix A : Plots and Summaries

Fig 1 Correlation Plot - Indication of Presence of Multicollinearity

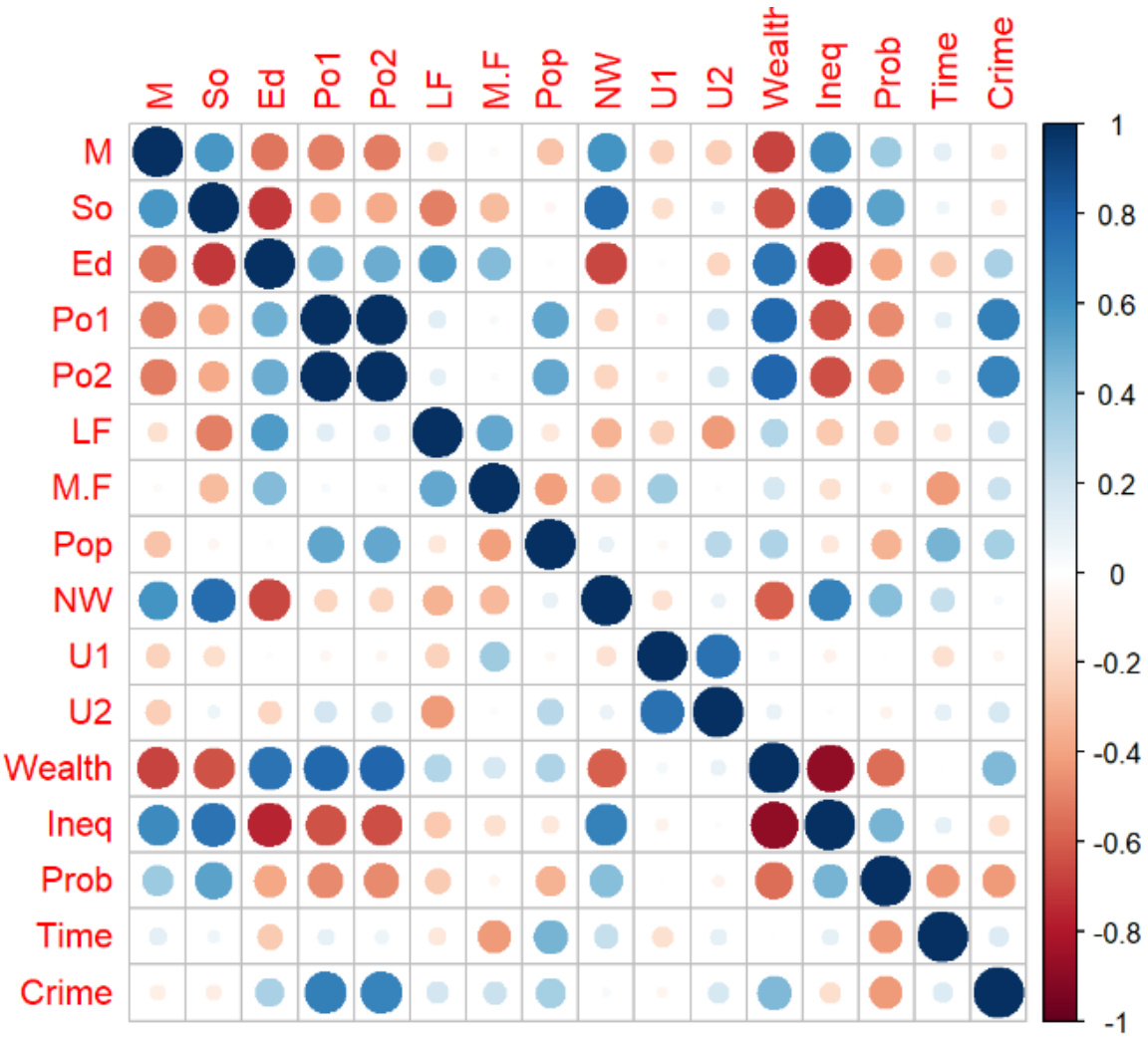

Fig 2 Comparison of Prediction Errors (Test RMSE)

The models numbers in the plot below are as follows :

1. Linear Regression
2. Forward Selection Method
3. Backward Elimination Method
4. Both ways selection method
5. Adjusted R-squared Based
6. AIC Based Selection
7. BIC Based Selection
8. Cp Mallow's Based Selection
9. Ridge Regression
10. Lasso Regression
11. Principal Components Regression

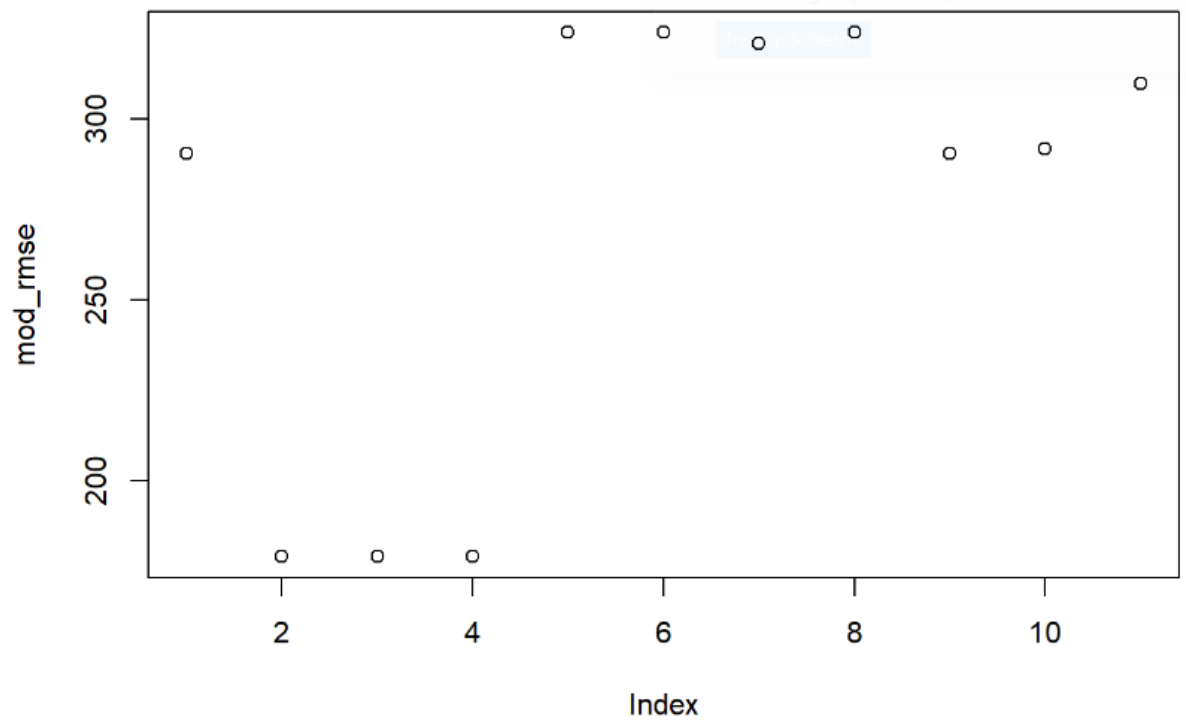

Fig 3 Regression Summary of Selected Model

```
##
## Call:
## lm(formula = Crime ~ So + Ed + Po1 + M.F + Ineq + Time, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -326.00 -153.14   30.74  108.54  264.02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7284.13    1359.38  -5.358 1.48e-05 ***
## So            184.41     116.50   1.583  0.12602
## Ed            180.70      48.73   3.708  0.00105 **
## Po1           125.59      15.37   8.172 1.59e-08 ***
## M.F            33.57      13.38   2.508  0.01901 *
## Ineq           65.39      18.83   3.473  0.00189 **
## Time           22.19       5.22   4.251  0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 175.6 on 25 degrees of freedom
## Multiple R-squared:  0.8422, Adjusted R-squared:  0.8044
## F-statistic: 22.24 on 6 and 25 DF,  p-value: 6.743e-09
```

Appendix B : References

1. https://www.jstor.org/stable/1831025