Group-6:

Ajinkya Phanse - amp660
Harsha Vardhan Tamatam - ht394
Omkar Uttam Ingle - oi64
Vaibhav Kumar - vk510

Analyzing Impact of Repeated Incarceration on Crime Rates In Communities

INTRODUCTION

The interplay between incarceration rates and crime has long been the subject of intense debate and empirical scrutiny. As public and policy interest in the phenomenon of mass incarceration — colloquially termed as 'prison cycling' — grows, it becomes imperative to dissect its efficacy and repercussions on crime rates within communities. This study embarks on an analytical journey, utilizing a linear regression framework to explore the relationship between repeated incarceration and crime rates, while also examining the influence of various socio-economic factors. Building on a rich dataset encompassing crime rates, prison cycling data, and census tract information from 2008 to 2010, this article seeks to unravel the complexities of crime dynamics in a nuanced socio-economic landscape. By challenging preconceived notions and testing prevalent hypotheses, the ensuing discourse contributes to the critical discourse on the viability of incarceration as a tool for crime deterrence and the socio-economic underpinnings of criminal behavior.

DATA & METHODS

The original dataset incorporates four Stata (.dta) files representing distinct geographic areas: Boston, MA; Newark, NJ; Trenton, NJ; and rural New Jersey. We have restricted our analysis to Boston region. Each observation in the data is indexed by the unique tract ID and year. In total there are 44 variables. After careful consideration we have chosen ten variables for analysis.

The study analyzes ten variables, integrating prison cycling data with census tract information, using unique eleven-digit tract identifiers. The dependent variable, TotalCrimeRate, combines standardized total violent and non-violent crimes by tract population. Independent variables include Cycle rate, Unemployment rate, Residential Mobility, Single-parent households, Median Income, Poverty Rate, and Population size. Data summary is provide in this table.

In explorative data analysis box plots are used to study the presence of skewness and outliers in data. Individual association between each independent variable and dependent variable is analyzed using scatter plots. To check correlation between independent variables correlation matrix (represent as heat map) is utilised.

2. MODEL DEVELOPMENT

The initial analysis involved fitting a base model with seven independent variables. According to this model only three variables cyclerate, unemployment rate, and population are significant. The R^2 and adjusted R^2 values are 0.33 and 0.29, which are relatively low.

2.1 Assumption Validation for Linear Regression

1. **Constant Variance of Residuals:**

   Upon fitting the base model, an assessment of the assumption of constant variance of residuals is conducted. The plot of residuals against fitted values exposes a concerning trend of increasing variance with fitted values, indicating the presence of heteroskedasticity. To address the issue of unstable variance, the Box Cox technique is employed. The log likelihood function is utilized, revealing maximization at lambda equal to 0, indicative of a logarithmic transformation. Post log transformation of the dependent variable and refitting the model, a stabilization of variance is observed (log transformed plot).

2. **Normality of Residuals:**

   The next assumption examined is the normality of residuals. This assessment is conducted using QQ plots and the Shapiro-Wilk test. The QQ plots display a distinct pattern suggestive of a heavy-tailed distribution of residuals. Furthermore, the P-value = 8.198e-05 from the Shapiro-Wilk test substantiates the deviation from normality.

3. **Autocorrelation In Residuals:**

   Examining the assumption of no autocorrelation in residuals, residuals are plotted against their lagged versions (residuals vs lagged residuals plot). A systematic pattern emerges, with observations concentrated in the left-bottom and top-right

quadrants. The Durbin-Watson test yields a p-value of 0.0012, indicating the presence of autocorrelation.

There are some serious violations of assumptions. Rather than jumping to the conclusion that the linear regression analysis is not suitable for the given data, these issues will be recalibrated after dealing with outliers and influential observations.

2.2 Detection and removal of Influential Points, Outliers

1. Removal of Influential Points and Outliers :

In the next phase of our analysis, we addressed unusual observations using Cook's distance, removing three influential points and two outliers through a sequential process. Following this data cleaning, we revisited assumptions related to linear regression. Notably, residuals normalized, the Shapiro-Wilk test for normality passed, and autocorrelation significantly decreased. With these improvements, our dataset is now deemed suitable for linear regression analysis.

2. Model Comparison post Data Cleaning and Transformation :

Following variable transformation and data cleaning, the resultant model emerges with notable improvements. In contrast to the baseline model, an additional variable has achieved significance. The R-squared value has shown a substantial increase, climbing from 0.33 to 0.54, while the adjusted R-squared has also experienced a notable rise, ascending from 0.29 to 0.51.

2.3 Detection and removal of Multicollinearity

Multicollinearity is assessed through the examination of the correlation matrix and the calculation of Variance Inflation Factors (VIF). The highest VIF observed is 4.34, attributed to the variable "poverty_rate," which remains below the commonly accepted threshold of 10. In an attempt to address multicollinearity, we experimented with a model excluding "poverty_rate," yet no substantial alterations were observed in the model. Consequently, no independent variables were removed.

2.4 Model Selection

For model selection purposes, a forward search approach based on three information criteria (AIC, BIC, and adjusted R-square) was employed. A recursive method was followed, incorporating a new independent variable at each step and re-fitting the model accordingly. The sequence in which parameters are incorporated into the model is determined by the significance of each parameter. According to AIC (AIC plot), the optimal number of variables is five, while according to BIC (BIC plot), it is either three or five (both having nearly equal values).Ultimately, the decision was made to proceed with five variables. This decision was influenced by the observation that the adjusted

R-square (adjusted R-square plot) shows improvement in the model with five variables compared to three variables. Additionally, the aim was to retain any extra information captured by the inclusion of the two additional variables, rather than discarding them.

2.5 Best Model

Summary (Best Model)

The best model included the following predictors: **cycle rate, the proportion of single-parent households, unemployment rate, residential mobility,** and **population size**(**Median-Income & Poverty-Rate** predictors are not considered as they doesn't have significant impact on estimating or explaining crime rate).

The regression model was statistically significant, $F(5, 145) = 32.54$, $p < .00001$, indicating that the model was able to distinguish the relative impact of the above stated predictors on the total crime rate. With an **Adjusted R^2 of 0.5125**, which suggests that approximately 52.87% of the variance in the total crime rate can be explained by the model.

The regression analysis demonstrated that cycle rate and unemployment rate are quite significant predictors and are positively associated with an increase in the total crime rate. The statistical confidence of pop & sharesingleparent is low and their coefficient impact on the estimator is also low. Residential mobility also showed a positive association with the total crime rate. These results suggest that along with prison cycling, socio-economic factors play a significant role in the variation of crime rates across different locales.

DISCUSSION & CONCLUSION

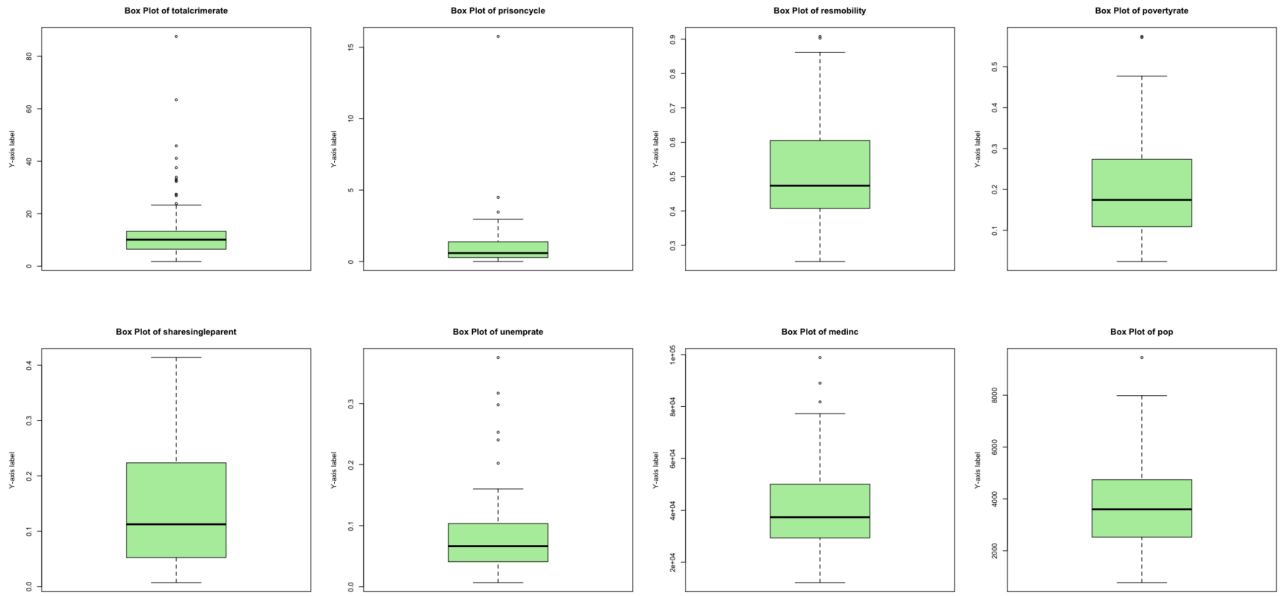
The implications of the findings from the linear regression model used to analyze prison cycling and various socio-economic factors are multifaceted and contribute significantly to the debate on mass incarceration and crime rates. Contrary to the commonly held belief that increased incarceration rates lead to a reduction in crime, our analysis indicates that prison cycling might actually have a positive correlation with the total crime rate in the long run. This suggests that policies aimed at reducing crime through mass incarceration may not have the intended effect and might even exacerbate the problem.

Furthermore, the notion that higher population density is a determinant of increased crime rates does not find strong support from our model. The population coefficient was found to be of minimal significance(and is inversely related to crime rates), which contradicts the examples often cited from densely populated areas like New York City and certain Californian cities. This finding challenges the perspective that population density is a substantial factor in crime rate fluctuations and suggests that policymakers should consider other variables when addressing crime prevention.

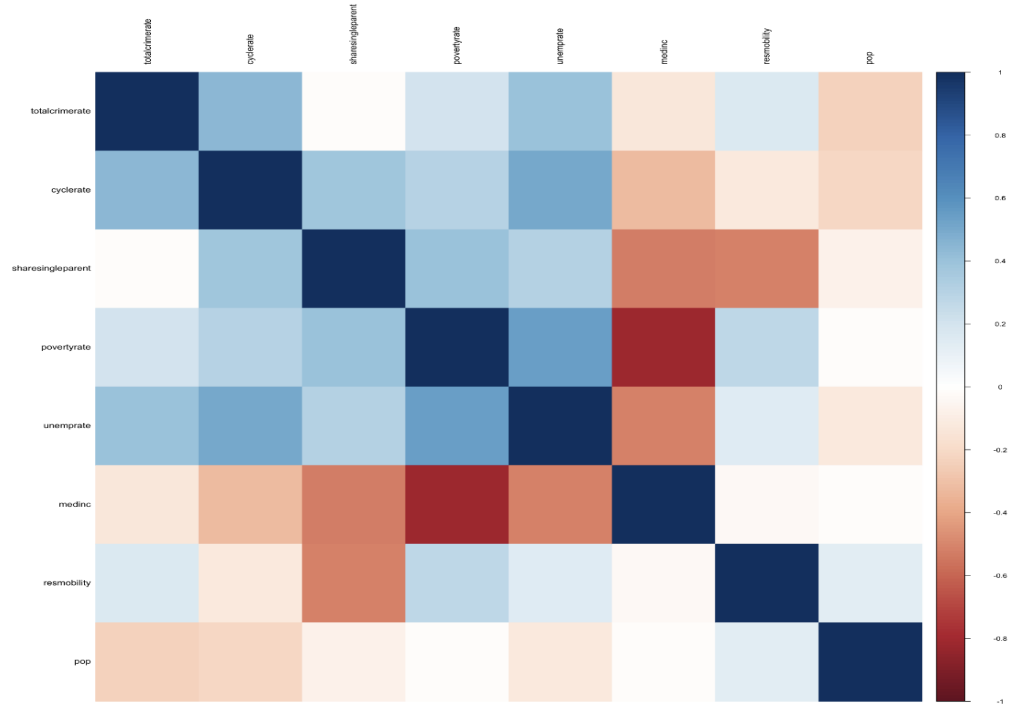
Socioeconomic factors are undeniably complex and interconnected. While median income and poverty rates might intuitively be considered as predictors of crime rates, our model indicates that they do not hold significant predictive power in this context. In contrast, unemployment rates and residential mobility emerge as noteworthy factors. This underscores the importance of economic stability and community cohesion in crime prevention strategies. Unemployment, which may reflect economic distress, and residential mobility, which might indicate social instability, are shown to be more closely associated with crime rates.

Appendix

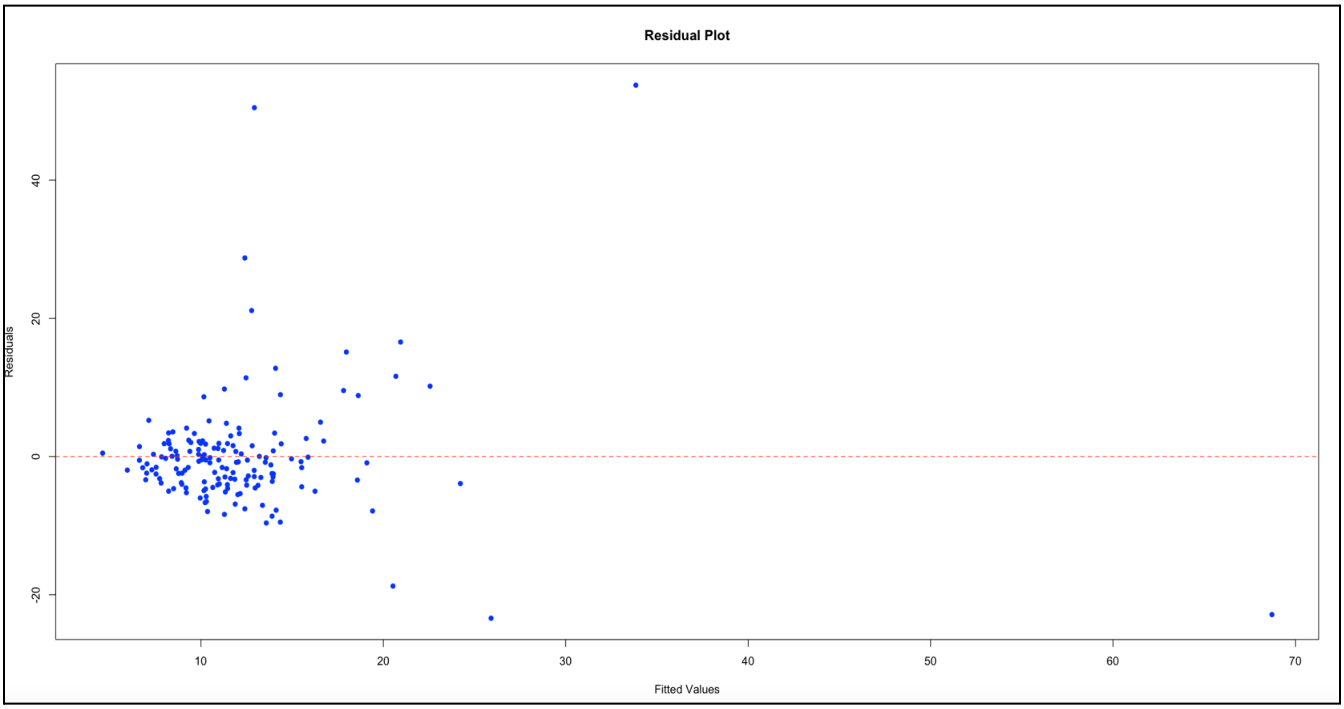
1. Box Plot (Predictors)



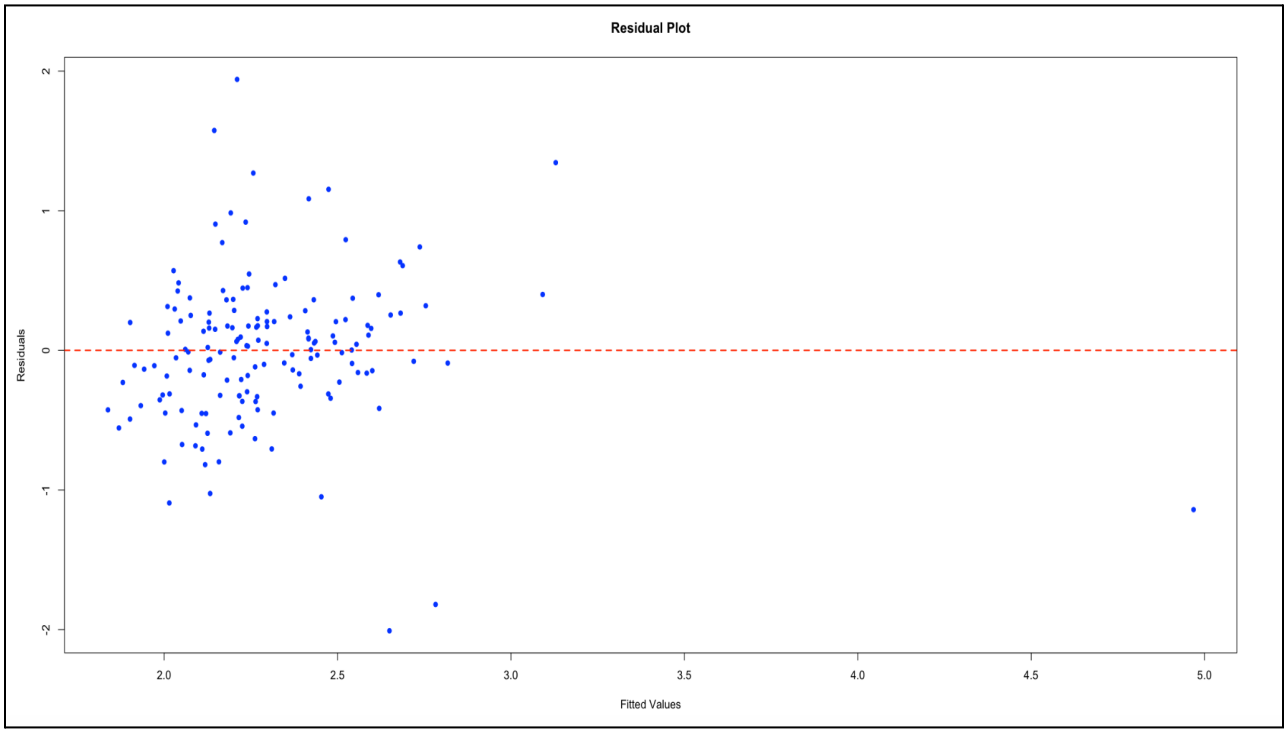
2. Correlation Heatmap



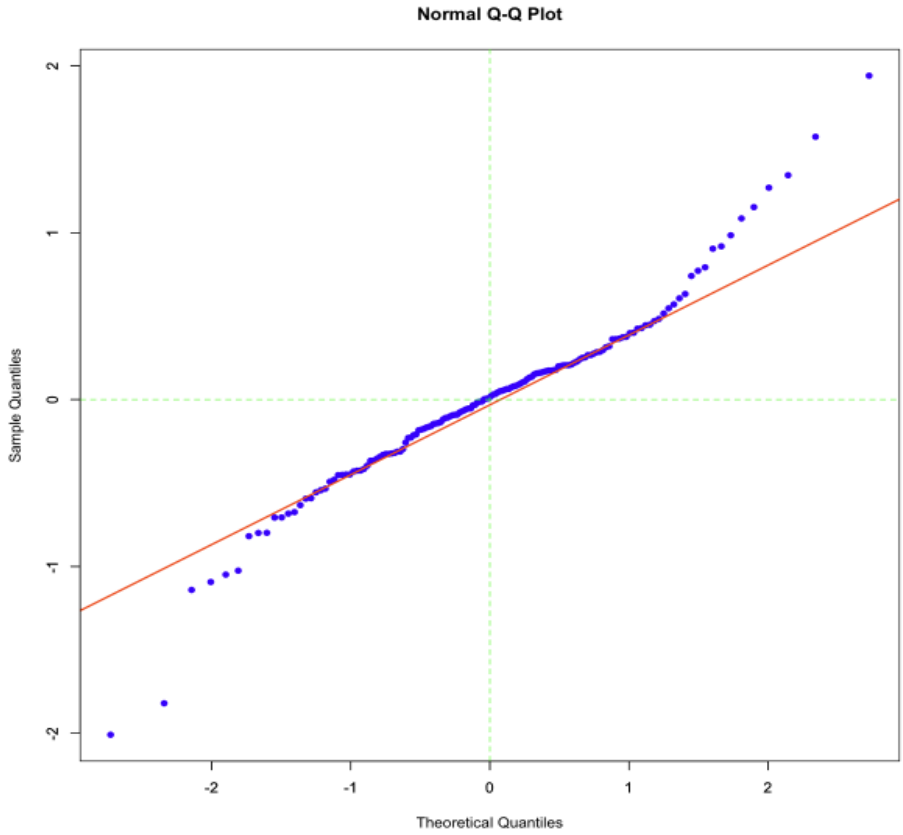
3. Residual plot (Base Model)



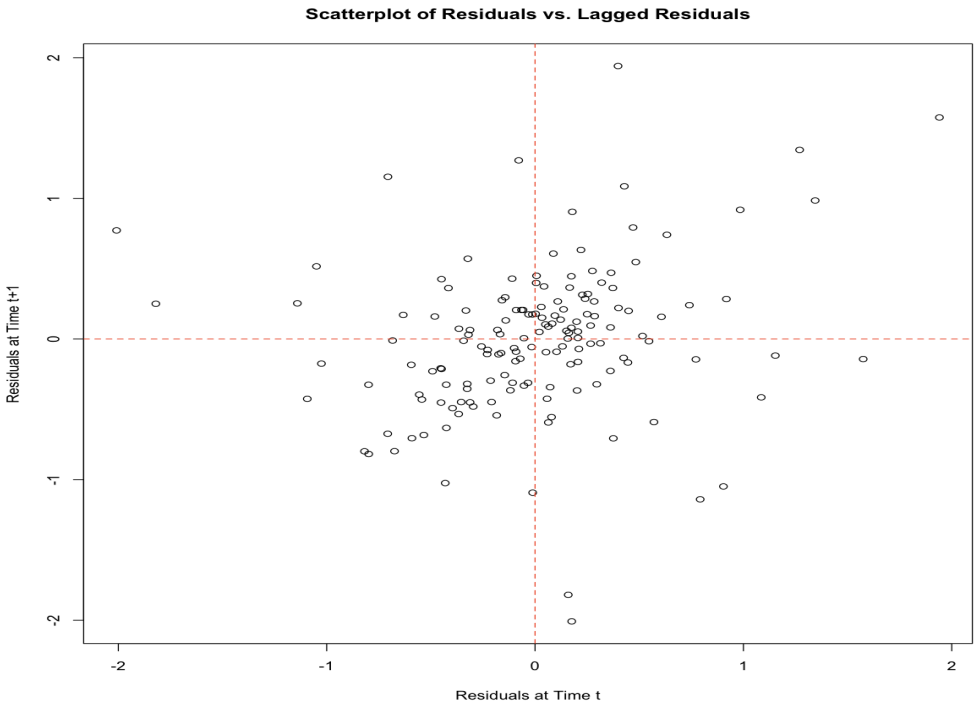
4. Residual plot (Log transformed Model)



5. Quantile-Quantile plot



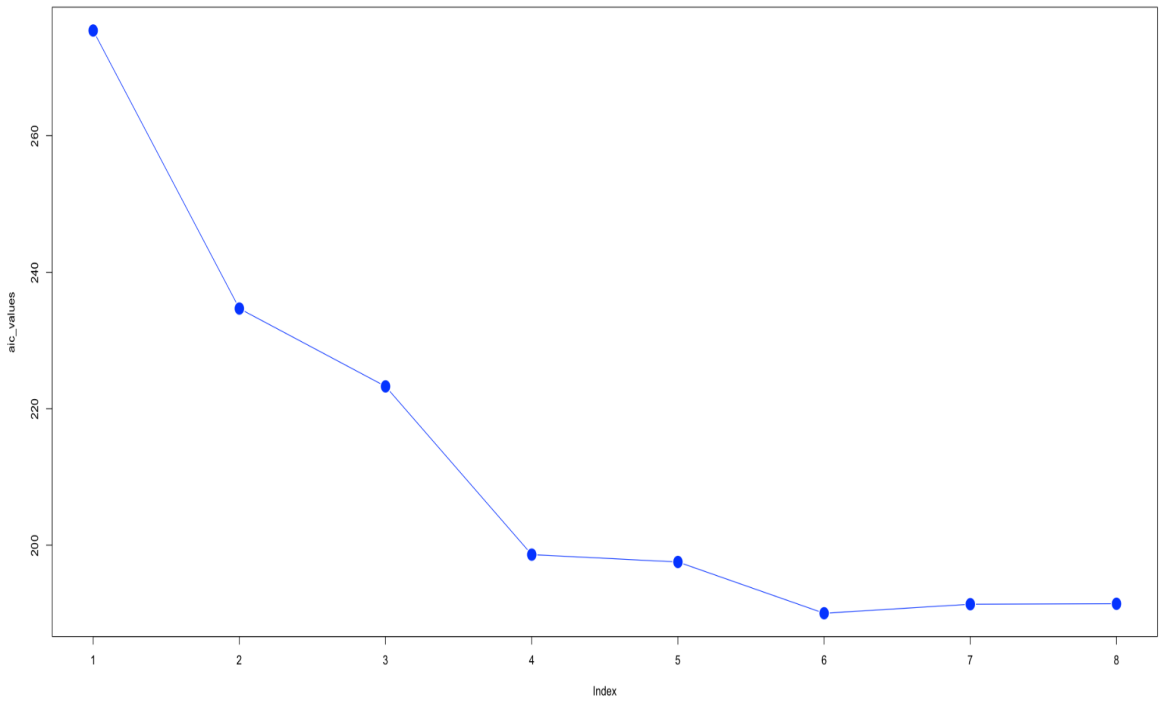
6. Residual vs Lagged Residuals plot



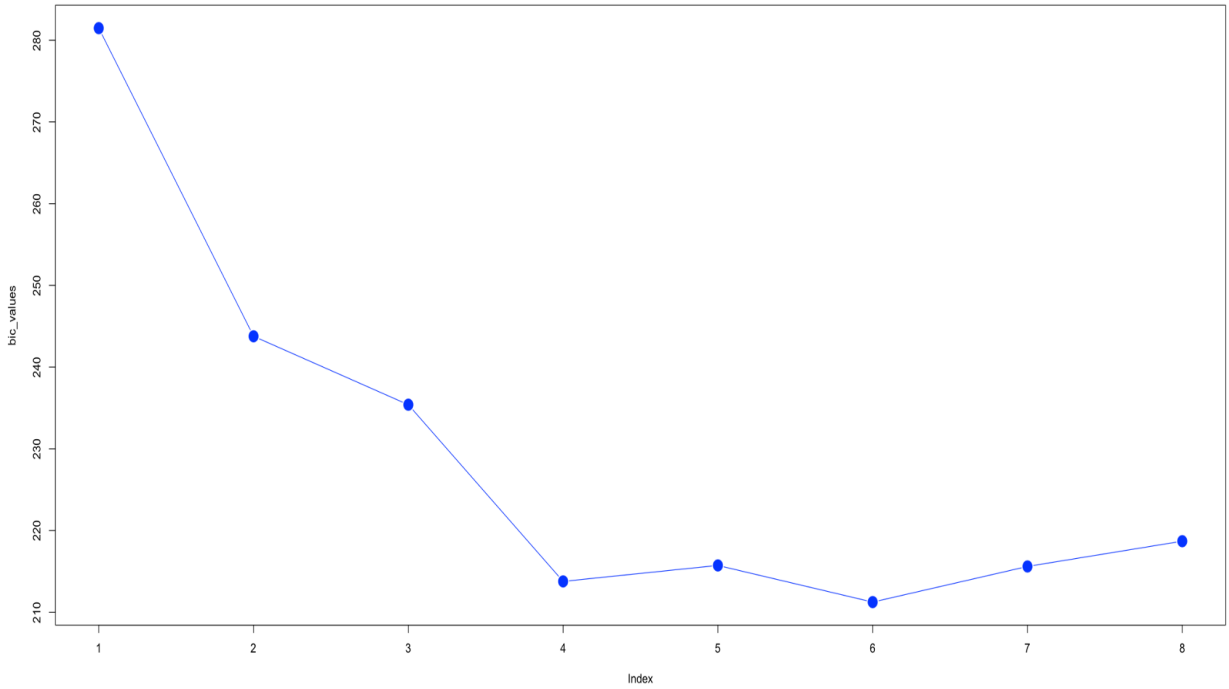
7. VIF Table

Variables	VIF values
cyclerate	2.537324
sharesingleparent	3.676932
povertyrate	**4.346610**
unemprate	1.787465
resmobility	2.249328
pop	1.096324
medinc	**3.673920**

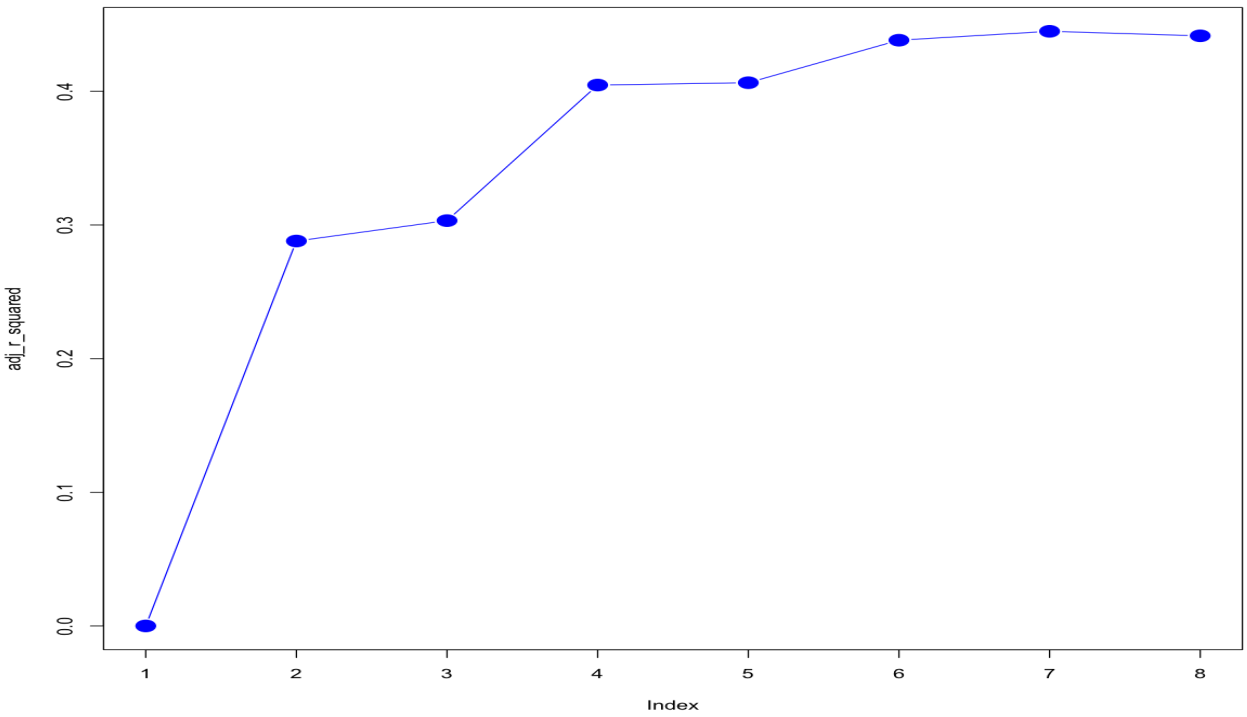
8. AIC Plot



9. BIC Plot



10. Adjusted R^2 Plot



11. Summary(Best Model)

```
Call:
lm(formula = totalcrimerate ~ cyclerate + sharesingleparent +
    unemprate + resmobility + pop, data = m6)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02278 -0.26258 -0.01283  0.25564  1.32604

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.541e+00  1.847e-01   8.344 5.14e-14 ***
cyclerate          4.025e-01  6.362e-02   6.326 2.94e-09 ***
sharesingleparent -1.245e+00  5.152e-01  -2.417  0.01688 *
unemprate          4.238e+00  8.328e-01   5.090 1.10e-06 ***
resmobility        8.022e-01  2.927e-01   2.740  0.00691 **
pop               -4.585e-05  2.130e-05  -2.153  0.03299 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3966 on 145 degrees of freedom
Multiple R-squared:  0.5287,    Adjusted R-squared:  0.5125
F-statistic: 32.54 on 5 and 145 DF,  p-value: < 2.2e-16
```

12. In Data Summary

Variable Name	Description
TotalCrimeRate	Total crime ('totalcrime') in a given quarter divided by the population size for a given tract ('pop') that has been divided by 1,000
CycleRate	Total cycling to and from correctional facilities ('cycle') in a given quarter divided by the population size for a given tract ('pop') that has been divided by 1,000
Census Tract identifiers	Eleven digits unique identifier assigned for every geolocation
Unemprate	Proportion (range: 0-1) of a given U.S. Census tract that is unemployed
Resmobility	Proportion (range: 0-1) of a given U.S. Census tract moved within the past five years
Sharesingleparent	Proportion (range: 0-1) of a given U.S. Census with households headed by a single parent
Medinc	Median income for a given tract corresponding with the U.S. Census
Povertyrate	Proportion (range: 0-1) of a given U.S. Census tract with reported family incomes
Pop	Population size for a given tract
Year	Year of crime event (2008-2010)

REFERENCES

1. Clear, Todd R., Frost, Natasha A., Carr, Michael, Dhondt, Geert, Braga, Anthony, and Warfield, Garrett A.R. Predicting Crime through Incarceration: The Impact of Prison Cycling on Crime in Communities in Boston, Massachusetts, Newark, New Jersey, Trenton, New Jersey, and Rural New Jersey, 2000-2010. Inter-university Consortium for Political and Social Research [distributor], 2017-03-22. https://doi.org/10.3886/ICPSR35014.v1

2. Kirk, Eileen M. "Community consequences of mass incarceration: sparking neighborhood social problems and violent crime." *Journal of Crime and Justice* 45.1 (2022): 103-119.

3. Clear, Todd R. Imprisoning Communities: How Mass Incarceration Makes Disadvantaged Neighborhoods Worse. New York, NY: Oxford University Press.

4. Carr, Michael, Dhondt, Geert Modeling the impact of incarceration on crime at the community level. American Society of Criminology 2013 Annual Meeting. Atlanta, GA.