Socioeconomic Factors Influencing Violent Crime Rates via Linear Regression Analysis

CHAYAPATR ARCHIWARANGUPROK

This study investigates the complex relationship between socioeconomic factors and violent crime rates in U.S. communities using the UCI Communities and Crime dataset. Initial analysis through linear regression revealed significant heteroscedasticity, indicating that the influence of socioeconomic factors varies systematically with community wealth levels. We addressed this through an income-based segmentation approach, which not only resolved the statistical issues but also unveiled distinct patterns of crime predictors across different economic strata. Our findings show that racial demographics dominate predictive power in low and middle-income communities (MI=0.3171 and MI=0.4188, respectively), while family structure emerges as the strongest predictor in high-income areas (MI=0.3030). The analysis also revealed unexpected patterns, with poverty indicators showing increasing importance in wealthy communities (MI=0.2376) and immigration-related factors becoming particularly relevant in middle-income areas (MI=0.2150). These results suggest that effective crime prevention strategies should be tailored to community characteristics rather than applying uniform solutions across diverse populations. The study contributes to both the methodological approach to crime analysis and practical policy development for community-specific crime prevention strategies.

1 Introduction

Understanding the determinants of violent crime remains a critical challenge in social policy and urban planning. While numerous studies have examined the relationship between socioeconomic factors and crime rates, the complex and varying nature of these relationships across different community types often goes unaddressed. This gap in understanding can lead to oversimplified or ineffective crime prevention strategies that fail to account for community-specific dynamics.

The UCI Communities and Crime[1] dataset provides a unique opportunity to examine these relationships in detail, combining socioeconomic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. With 122 predictive features ranging from demographic composition to law enforcement statistics, this dataset enables a comprehensive analysis of factors influencing violent crime rates across diverse communities.

Through this analysis, we aim to contribute to both the theoretical understanding of crime determinants and the practical development of community-specific crime prevention strategies. Our findings challenge several conventional assumptions about crime predictors and suggest the need for more nuanced, segment-specific approaches to crime prevention and policy development.

2 Dataset

The Communities and Crime dataset combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The dataset originally contained 127 features, with the target variable being per capita violent crimes (including murder, rape, robbery, and assault). The features can be categorized into several groups:

- Demographic indicators (e.g., population, household size, racial percentages)
- Age distribution metrics (e.g., percentage of population aged 12-21, 16-24, 65 and up)
- Economic indicators (e.g., median income, percentage with wage income, per capita income by race)
- Education metrics (e.g., percentage with less than 9th grade, percentage with BS or higher)
- Employment statistics (e.g., percentage unemployed, percentage in manufacturing)

[1] https://archive.ics.uci.edu/dataset/183/communities+and+crime

Author's Contact Information: Chayapatr Archiwaranguprok, pub@from.pub.

- Family structure indicators (e.g., percentage of two-parent families, divorce rates)
- Housing characteristics (e.g., persons per household, housing occupancy rates)
- Law enforcement statistics (e.g., police personnel data, budget allocation)

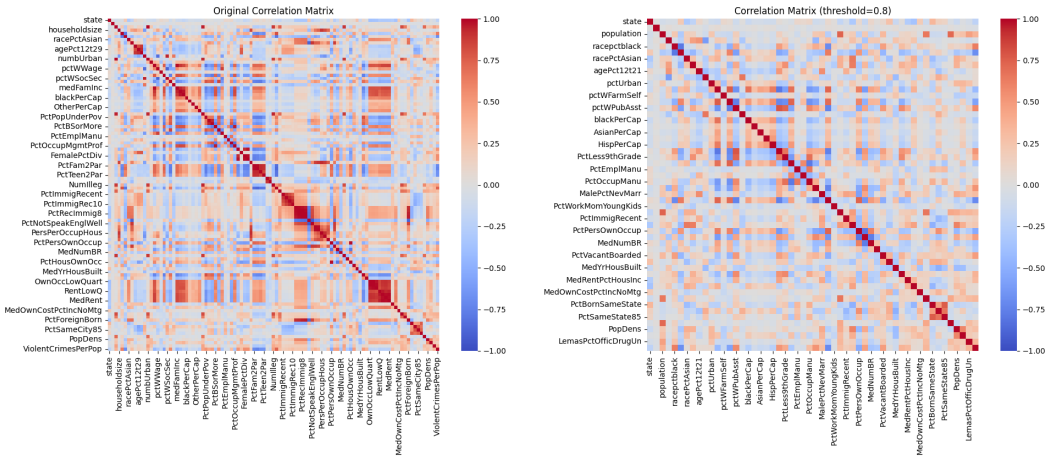

Fig. 1. Feature Correlation Matrix Before and After Selection

With 127 features, the dataset presented challenges of dimensionality and multicollinearity. To address this, we perform a feature selection process by creating a Pearson correlation coefficient matrix and then removing highly correlated features (threshold=.8). In effect, we reduced the feature set to 53 significant predictors. The correlation analysis (Figure 3) shows the original feature space and the reduced feature set after applying the correlation threshold, demonstrating the effectiveness of our feature selection process in reducing redundancy while maintaining important predictive relationships.

3 Linear Regression

3.1 Model Performance

The initial Ordinary Least Squares (OLS) regression analysis achieved moderate predictive power with a training R^2 of .673 and test R^2 of .636, indicating that approximately 64% of the variance in violent crime rates can be explained by the model. This suggests a reasonable but not strong predictive capability, pointing to the complex nature of crime prediction. This range of numbers is commonly encountered in the field of sociology.

3.2 Feature Importance

Analysis of feature importance revealed several key factors strongly associated with violent crime rates:

- Family structure was the most influential predictor, with the percentage of two-parent families (PctFam2Par) having the highest importance score of .385
- Homelessness indicator (NumStreet) emerged as the second most important feature (.243)
- Racial demographics, specifically the percentage of the Black population (racePctBlack), showed significant importance (.227)
- Household characteristics (householdsize: .193) and age distribution (agePct65up: .180) also emerged as notable predictors

- Economic indicators such as investment income percentage (pctWInvInc: .147) and retirement income percentage (pctWRetire: .123) showed moderate importance

3.3 Model Diagnostics and Assumption Testing

Diagnostic analysis of the model revealed significant violations of key regression assumptions. The scatter plot of actual versus predicted values shows increased scatter and underprediction at higher crime rates while maintaining better accuracy for communities with lower crime rates. The residuals plot exhibits a distinct fan-shaped pattern, indicating heteroscedasticity that increases with fitted values.

These visual observations are confirmed by formal statistical tests:

- The Shapiro-Wilk test (W=.960, P<.001) indicates the non-normal distribution of residuals, particularly visible in the Q-Q plot's tail deviations
- The Breusch-Pagan test (LM=95.36, P<.001) confirms severe heteroscedasticity in the residuals

These violations suggest that a simple linear regression approach may not adequately capture the underlying relationships in the data, necessitating model refinements to address both the non-normality of residuals and heteroscedasticity.

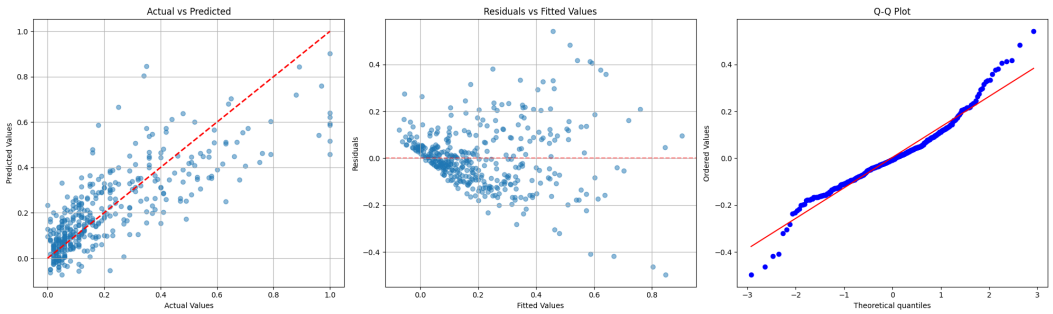

Fig. 2. Diagnostic Plots for OLS Linear Regression

4 Model Refinement

4.1 Yeo-Johnson Transformation

To address the normality concerns, we applied the Yeo-Johnson transformation to the target variable. After transformation, the Shapiro-Wilk test (P=.11, P>.05) indicated improved normality of residuals, as also evidenced by the Q-Q plot alignment. However, the Breusch-Pagan test (P=.0048) revealed persistent heteroscedasticity in the residuals, suggesting the need for additional model refinements. The residuals vs fitted values plot confirmed this finding, showing a distinctive fan shape pattern indicative of non-constant variance.

4.2 Alternative Approaches

Several methods were attempted to address the persistent heteroscedasticity:

- Weighted Least Squares (WLS) using inverse squared residuals
- Huber regression with varying epsilon values (1.1 to 1.75)
- RANSAC regressor for outlier robustness
- A combined approach using Huber regression with weights

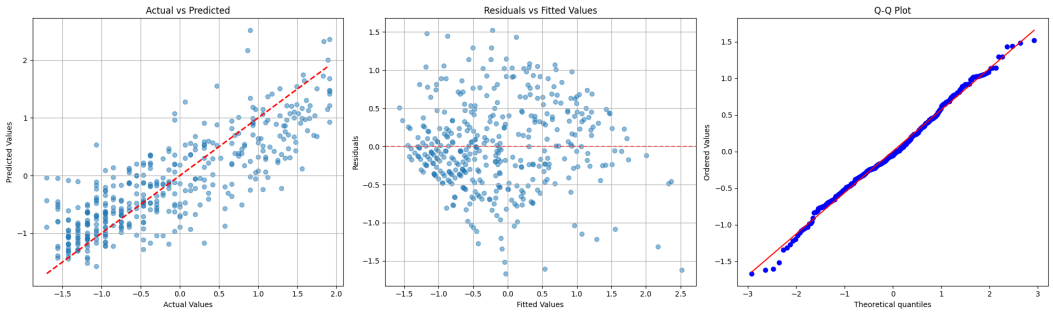

Fig. 3. Diagnostic Plots for Linear Regression after Yeo-Johnson transformation

However, all these methods yielded Breusch-Pagan P-values below .05, indicating unresolved heteroscedasticity. This consistent violation across different approaches suggested a more fundamental structure in the data variance, leading us to explore factor-based segmentation as an alternative solution.

5 Factor-based Segmentation Analysis

5.1 Segmentation

We analyzed the relationship between feature values and residual variance through a systematic residual analysis. Each feature's contribution to heteroscedasticity was quantified by measuring the correlation between its values and the squared residuals from the initial model. Visual inspection of residual plots against individual features (Figure 4) revealed varying patterns of heteroscedasticity across different factors. The analysis showed that economic indicators were particularly problematic, with median income (medIncome) demonstrating the strongest relationship with residual variance (correlation=.262). The top contributing factors to heteroscedasticity are listed in Table 1.

Category	Indicator	Value
Economic Factors	Median income	.262
	Investment income percentage	.226
	Public assistance percentage	.189
Social Structure Indicators	Male divorce rate	.222
	Education level (BS or higher)	.206
	Two-parent family percentage	.204
Housing Characteristics	Long-term vacant properties	.195
	Housing occupancy rate	.188
	Building age	.170

Table 1. Socioeconomic and Housing Indicators Analysis

The strong relationship between economic factors and residual variance, particularly median income, suggests that the impact of various predictors on violent crime rates varies systematically across different income levels. This finding motivated the use of income-based segmentation as a strategy to address the heteroscedasticity issue.

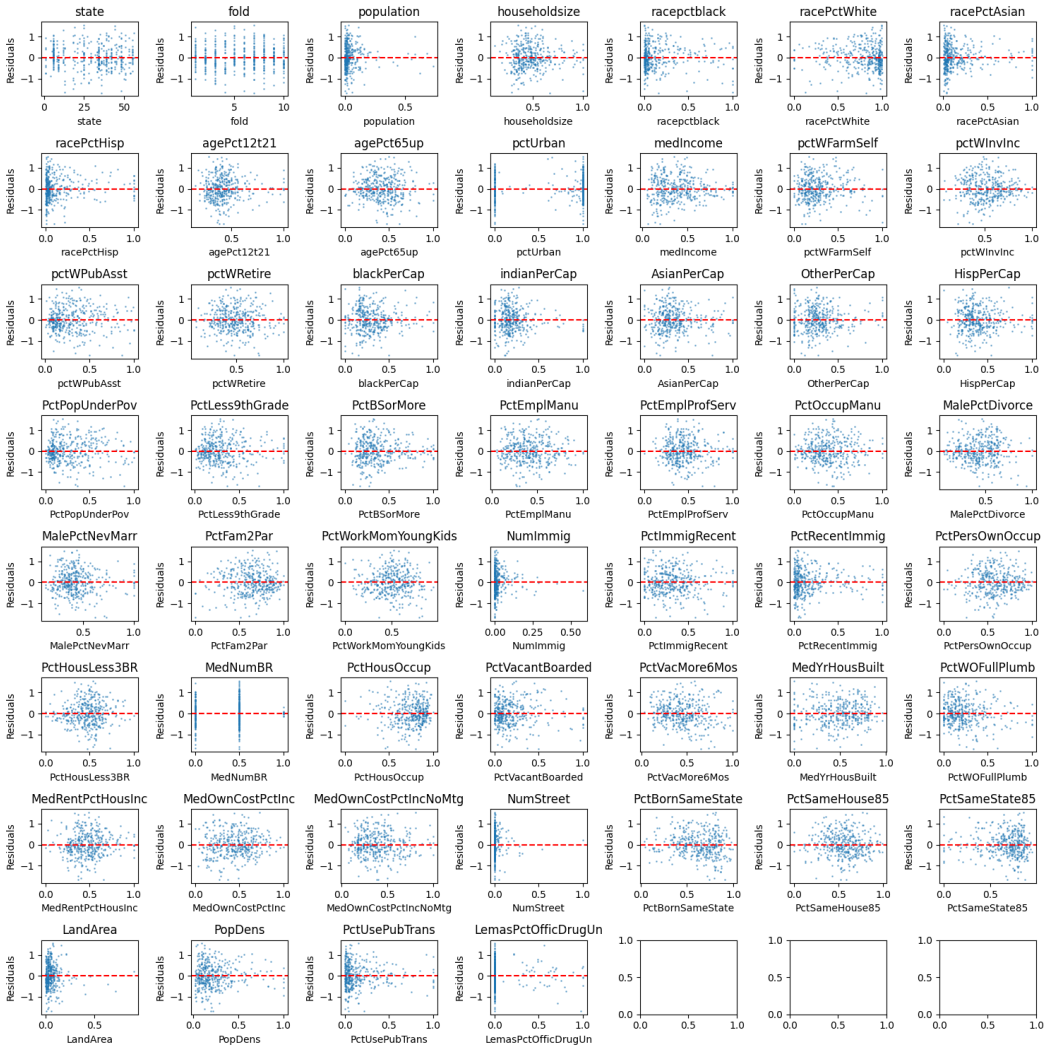

Fig. 4. Residual Plots Against Individual Features

5.2 Segmented Linear Regression

The segmented model demonstrates several key improvements in terms of model diagnostics despite a marginal decrease in overall predictive performance from R^2 .679 to .653 (-2.66%). Most notably, the model successfully addresses the heteroscedasticity issues present in the original model, with the Breusch-Pagan P-value improving from .000201 to .068330, exceeding the conventional .05 significance threshold. The segmentation across income levels reveals distinct patterns in model performance and diagnostic metrics. **The low-income segment** (.000-.230) encompasses 108 observations and achieves an R^2 of .508, demonstrating excellent normality (SW P=.840) and stable variance (BP P=.520). **The middle-income segment** (.230-.420) performs best with an R^2 of .685 across 131 observations, maintaining both normality (SW P=.630) and homoscedasticity (BP P=.401). **The high-income segment** (.420-1.000), covering 160 observations, achieves an R^2 of .605 but

shows some deviation from normality (SW P=.003) while maintaining homoscedastic residuals (BP P=.371).

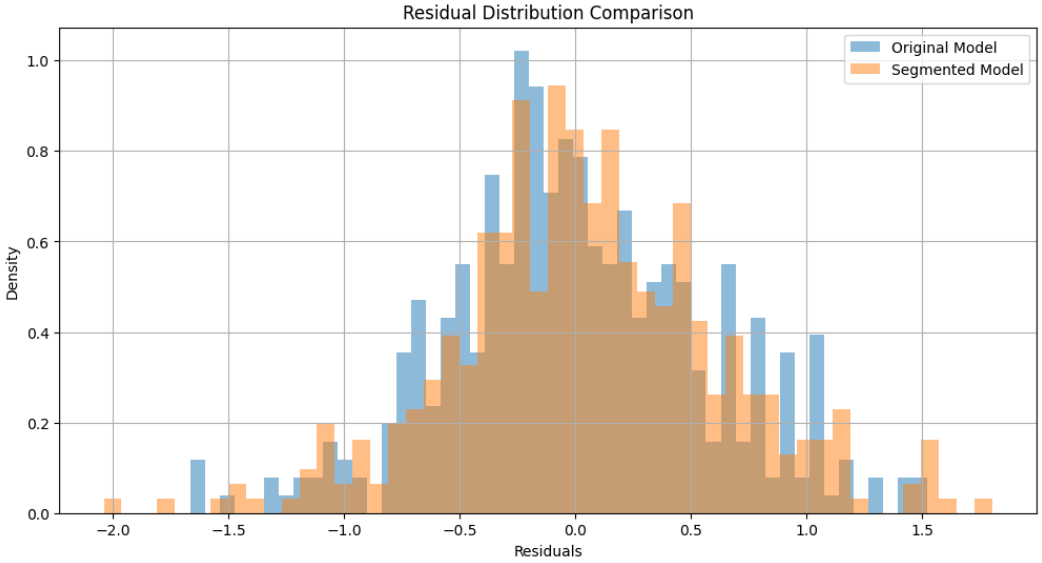

Fig. 5. Residual distribution comparison between original and segmented model

The overlaid residual distributions in Figure 5 illustrate the improved symmetry and reduced tail thickness in the segmented model compared to the original model. While the overall model's Shapiro-Wilk test indicates a slight departure from normality (P=.047), the segment-wise analysis reveals the high-income segment primarily drives this, as both low and middle-income segments show strong adherence to normality assumptions. This heterogeneous performance across income segments, particularly the strong normality and homoscedasticity in low and middle-income brackets coupled with distinct behavior in high-income areas, suggests that further refinement of segment-based modeling approaches could yield more nuanced and reliable predictions for specific socioeconomic contexts, potentially through the exploration of adaptive modeling strategies or segment-specific feature selection.

6 Results and Discussion

The segmentation analysis revealed distinct patterns of crime predictors across different income levels, providing nuanced insights into how socioeconomic factors influence crime rates differently across community segments. Due to the regression coefficients being affected by the Yeo-Johnson transformation, we employ Mutual Information (MI) scores to analyze key patterns in feature importance.

6.1 Evolution of Feature Importance Across Segments

- **Racial Demographics** emerged as a crucial factor, particularly in lower and middle-income segments. The percentage of the white population (racePctWhite) showed the strongest predictive power in both low-income (MI=0.3171) and middle-income segments (MI=0.4188), while its importance decreased significantly in high-income communities (MI=0.1628). This pattern was similarly reflected in the percentage of the Black population (racepctblack),

which showed high importance in low-income areas (*MI*=0.2925) but diminished influence in higher-income segments.

- **Family Structure**, measured by the percentage of two-parent families (PctFam2Par), demonstrated remarkable consistency across all segments while showing varying degrees of importance. It ranked third in low-income communities (*MI*=0.2795), second in middle-income areas (*MI*=0.2604), and emerged as the most important predictor in high-income segments (*MI*=0.3030). This consistent presence suggests that family structure is a universal factor in crime prediction, though its relative importance varies with community wealth levels.
- **Immigration and Poverty** factors showed distinct patterns across segments. Immigration-related features (NumImmig, PctRecentImmig) became particularly relevant in middle-income communities (*MI*=0.2150, 0.1869 respectively) while maintaining moderate importance in high-income areas. Poverty indicators (PctPopUnderPov) showed increasing importance with community wealth, becoming the second most important predictor in high-income segments (*MI*=0.2376).
- **State-level variation** (state) demonstrated decreasing importance across wealth segments, from *MI*=0.1907 in low-income to *MI*=0.1637 in high-income communities, suggesting that geographic factors may have less influence on crime rates in more affluent areas.

Rank	Feature	MI Score by Segment		
		Low	Middle	High
1	Race (White)	**0.3171 (1)**	**0.4188 (1)**	-
2	Race (Black)	0.2925 (2)	0.2024 (5)	-
3	Two-Parent Families	0.2795 (3)	0.2604 (2)	**0.3030 (1)**
4	State	0.1907 (4)	0.2501 (3)	0.1637 (9)
5	Investment Income	0.1669 (5)	-	0.1934 (6)
6	Male Divorce Rate	0.1548 (6)	-	0.1996 (5)
7	Less 9th Grade Education	0.1170 (7)	-	-
8	Public Assistance	0.1046 (8)	0.1933 (7)	0.2163 (3)
9	Population	0.1011 (9)	0.1806 (10)	-
10	Drug Unit Officers	0.0981 (10)	-	-
11	Immigration Count	-	0.2150 (4)	0.1708 (8)
12	Population Below Poverty	-	0.1934 (6)	*0.2376 (2)*
13	Race (Hispanic)	-	0.1906 (8)	0.2023 (4)
14	Recent Immigration	-	0.1869 (9)	-
15	Home Ownership	-	-	0.1822 (7)

Table 2. Feature Importance Comparison Across Income Segments (with rank order in parentheses)

6.2 Policy Implications

These findings suggest the need for tailored approaches to crime prevention across different community segments. **Low-Income Communities** should focus on addressing racial disparities and strengthening family structures, as these emerged as the strongest predictors. The high importance of law enforcement factors (LemasPctOfficDrugUn, *MI*=0.0981) suggests that targeted policing strategies may be particularly effective in these areas. **Middle-Income Communities** require a more diverse approach, with particular attention to immigration integration and community cohesion programs. The high predictive power of both racial demographics and immigration factors

suggests the importance of inclusive community development strategies. **High-Income Communities** should prioritize family support systems and poverty prevention, given the high importance of family structure and poverty indicators. The significance of public assistance (pctWPubAsst, MI=0.2163) suggests that maintaining strong social safety nets remains important even in wealthy areas.

6.3 Limitations and Future Research

Several limitations warrant consideration in interpreting these results. The temporal nature of the data (1990s) may not reflect current socioeconomic patterns, and geographic variations might influence the relationships observed. The interaction between segments might need further investigation to understand spillover effects and community interdependencies. The segmented approach demonstrated in this study provides a framework for understanding how different factors contribute to crime rates across economic strata, suggesting that crime prevention strategies should be tailored to community characteristics rather than applying uniform solutions across diverse populations. Future research should focus on longitudinal studies to track how these relationships evolve over time, investigation of inter-segment dynamics, and analysis of policy intervention effectiveness across different segments. Understanding these temporal and spatial dynamics could provide valuable insights for more effective, targeted crime prevention strategies.

7 Conclusion

This study analyzed the relationship between socioeconomic factors and violent crime rates across U.S. communities using the UCI Communities and Crime dataset, revealing distinct patterns through income-based segmentation and mutual information analysis. Our findings showed that racial demographics dominated predictive power in low and middle-income communities (MI=0.3171 and MI=0.4188, respectively), while their influence diminished in high-income areas. Family structure maintained consistent importance across all segments, ultimately emerging as the strongest predictor in high-income communities (MI=0.3030), while poverty indicators showed increasing predictive power with community wealth (MI=0.2376 in high-income areas). Immigration-related features proved particularly relevant in middle-income communities (MI=0.2150), challenging conventional narratives about crime determinants. These segment-specific patterns suggest that effective crime prevention strategies should be tailored to community characteristics rather than applying uniform solutions, with racial equity initiatives potentially more crucial in lower-income areas, while family support systems and poverty prevention might yield better results in high-income communities. Future research should validate these findings with contemporary data and explore the interaction effects between immigration and economic factors, particularly in middle-income communities, to develop more effective, targeted crime prevention strategies.