Understanding Racial Disparities and Systemic Inequities in Fatal Police Shootings Across the United States*

Black and Hispanic Individuals Face Higher Odds of Being Shot and Killed by Police, Highlighting Structural Inequities in Policing Practices

Jingchuan Xu

December 11, 2024

This study examines fatal police shootings in the U.S. by linking incident data from The Guardian with socioeconomic data from the American Community Survey. A generalized linear model reveals significant racial disparities, with Black and Hispanic individuals facing higher odds of being shot and killed by police compared to White individuals, even after adjusting for community factors. These findings emphasize the role of systemic inequities in policing and the need for reforms to address racial disparities in the use of lethal force.

1 Introduction

The association of police-involved fatal shootings and community-level socioeconomic factors has become a pivotal area of research, revealing systemic inequities and law enforcement practices. Fatal police shootings frequently expose structural disparities, disproportionately impacting marginalized communities and raising critical questions about equity and accountability (Zare et al. 2022). Prior research has highlighted race, poverty, and neighborhood characteristics as significant contributors to these outcomes, yet these factors are often studied in isolation or with limited integration of individual and community-level data (Nix and Shjarback 2021; Henderson et al. 2024). This study seeks to address this gap by linking detailed incident-level data on police shootings from The Guardian with tract-level socioeconomic data from the American Community Survey (ACS). By examining how individual factors, such as race and armed status, interact with community attributes like poverty and

*Code and data are available at: https://github.com/Jonathan666Charlie/StoryTelling.git.

unemployment, this analysis investigates the systemic and structural dimensions underlying fatal police shootings.

The primary estimand of this study is the conditional probability of being fatally shot by police, given specific individual attributes (e.g., race, armed status) and community-level predictors (e.g., poverty rate, unemployment level). This outcome, modeled as a binary variable (shot and killed vs. not shot), is operationalized using a frequentist generalized linear model (GLM) with a logistic link function. The GLM framework allows for the estimation of odds ratios to quantify the relative effects of predictors on the likelihood of a fatal shooting. For example, the model evaluates how the odds of being shot differ across racial groups while holding community-level variables constant. This approach not only quantifies disparities but also enables an exploration of the relationship between individual and structural factors, allowing us to understand systemic patterns in police use of force.

The findings reveal persistent racial disparities in fatal police shootings, with Black and Hispanic individuals disproportionately affected compared to White individuals, even after adjusting for community-level factors. Armed status, while statistically significant, exhibits a smaller effect size, indicating that whether an individual was armed is less predictive than systemic and structural inequities. These results underscore the centrality of race in predicting the likelihood of being shot and killed by police, suggesting that immediate factors like armed status are insufficient to explain these disparities. The analysis highlights the importance of addressing broader societal and institutional inequities to reduce the disproportionate use of lethal force against marginalized groups.

This paper is structured as follows. Section 2 provides a comprehensive description of the data integration process, including the alignment of police shooting incidents and socioeconomic variables from The Guardian and the ACS (FiveThirtyEight 2015a). Section 3 outlines the GLM framework used for the analysis, discussing key predictors, model assumptions, and diagnostic checks. Section 4 presents the findings, emphasizing the significant roles of race and armed status in predicting fatal police shootings and assessing the model's performance. Finally, Section 5 contextualizes the results within broader debates on policing equity, addresses the limitations of the study, and offers recommendations for future research to deepen our understanding of systemic factors influencing fatal police shootings.

2 Data

2.1 Overview

For this analysis, we used the `R` programming language (R Core Team 2023) to examine data on police killings in the United States, as reported by the Guardian and supplemented with tract-level census data from the American Community Survey (ACS) (FiveThirtyEight 2015a). The dataset provides detailed information about each incident, including demographic and geographic characteristics, as well as socioeconomic variables derived from ACS tables.

Key variables analyzed include the demographic details of the deceased (e.g., race/ethnicity, age, gender), the geographic context (e.g., census tract, income levels, poverty rates), and law enforcement agency involvement (FiveThirtyEight 2015b). This comprehensive dataset enabled an in-depth exploration of trends and disparities in police killings across various regions and demographic groups.

A range of `R` packages was used for efficient data processing, modeling, and validation. The `tidyverse` suite streamlined data cleaning and transformation (Wickham et al. 2019), while `here` ensured consistent file path management (Müller 2020). `janitor` handled missing values and standardized variables (Firke 2023), and `modelsummary` generated clear regression summaries (Arel-Bundock 2022). `arrow` optimized large dataset handling (Richardson et al. 2024), and `testthat` ensured data integrity (Wickham 2011). For evaluation, `pROC` calculated ROC-AUC to assess model performance (Robin et al. 2011), and `plumber` enabled API deployment of the model for real-time predictions (Schloerke and Allen 2024). This analysis adhered to best practices in workflow organization and data science methodology (Alexander 2023), ensuring seamless integration of data manipulation, modeling, and visualization. This approach enabled a comprehensive investigation of police killings, contextualized with tract-level socioeconomic data.

2.2 Data source

The dataset for this analysis combines police killing data from the Guardian with tract-level socioeconomic data from the ACS, unveiling the relationshp of law enforcement actions and community characteristics (FiveThirtyEight 2015a, 2015b). A meticulous data cleaning process ensured the dataset's readiness for analysis. Missing or incomplete entries were examined and addressed, with invalid geographic or demographic records excluded. Census tract identifiers were cross-verified to ensure alignment with ACS data, enabling the inclusion of variables such as poverty rates, unemployment levels, and income distribution. Additional derived variables, such as the proportion of the population by race/ethnicity and education level, were calculated to show more community demographics. This cleaning process ensured that the dataset was accurate, comprehensive, and capable of supporting robust statistical analysis. The integration of police incident data with socioeconomic indicators allowed for an exploration of disparities and systemic patterns, helping policymakers, researchers, and advocates equity and justice.

2.3 Measurement

The dataset on police killings in 2015 serves as a valuable resource for examining the intersection of law enforcement actions and community socioeconomic characteristics. Compiled from the Guardian's database, which records critical details of each incident—including the name, age, race/ethnicity, and gender of the deceased, along with the date, location, and cause of death—the data offers a granular look at these tragic events. This information is further enriched with tract-level data from the American Community Survey (ACS), which provides demographic and socioeconomic variables such as poverty rates, income distribution, and educational attainment. By combining these sources, the dataset shows systemic patterns and disparities in police killings across different regions.

The Guardian's records were meticulously geocoded to census tracts using latitude and longitude data, ensuring alignment with ACS estimates from the 2015 5-year release. These estimates include variables derived from key tables, such as demographics (S0601), income and poverty (S1901), and employment and education (S1701). Calculated metrics, such as income quintiles and ratios, provide additional layers of insight into the socioeconomic context of affected communities. Standardized identifiers, like GEOIDs, allow integration with other census datasets, ensuring that the information is both robust and compatible with broader analyses. This linkage offers a structured view of how geographic and demographic characteristics intersect with patterns in police actions, enabling comparisons across various social and economic dimensions.

While the dataset provides a detailed and standardized account of police killings, it is not without limitations. Incidents may be underreported or unevenly documented, and the ACS data, being based on estimates, may not fully capture the dynamic realities of rapidly evolving neighborhoods. Moreover, the dataset lacks qualitative dimensions that could provide a fuller context, such as community narratives or local law enforcement policies. These limitations highlight the need for careful interpretation and the potential value of integrating additional data sources in future research. Despite these challenges, the dataset remains a critical tool for analyzing disparities in police killings and their connection to broader structural inequalities.

2.4 Outcome variable

The outcome variable in this analysis is shot, indicating whether the deceased individual was fatally shot by law enforcement. As a binary variable, it captures whether a police shooting occurred, with a value of 1 representing a fatal shooting and 0 representing an alternative cause of death. This variable serves as the key measure for understanding the prevalence and context of fatal police shootings, enabling an examination of how specific demographic and situational factors influence the likelihood of this outcome.

2.5 Predictor variables

2.5.1 Race ethnicity

One critical predictor variable is race/ethnicity, which categorizes the deceased individual's racial or ethnic background, as documented in the Guardian's dataset. This variable captures categories such as White, Black, Hispanic, Asian, and Native American, among others. By examining race/ethnicity, the analysis can explore disparities in fatal police shootings across different racial groups, shedding light on potential systemic biases or inequities. This variable is particularly crucial for investigating racial dynamics in law enforcement actions and understanding how these dynamics intersect with broader socioeconomic and structural inequalities.

2.5.2 Armed status

Another key predictor variable is armed status, which specifies whether the deceased individual was in possession of a weapon at the time of the incident. This variable includes categories such as firearm, knife, or no weapon, among others. It provides critical context for understanding the circumstances surrounding police shootings, as the presence or absence of a weapon often plays a central role in determining the perceived threat level and subsequent use of force by law enforcement. Analyzing this variable allows for an assessment of how weapon possession correlates with the likelihood of fatal outcomes and whether it influences disparities observed across other predictors like race/ethnicity.

2.6 Relationships among key variables

Figure 1 illustrates the proportion of individuals shot by police ("Yes" or "No") across armed and unarmed statuses. The y-axis represents the proportion, scaled from 0 to 1, while the x-axis shows the armed status categories ("No" and "Yes"). A larger proportion of "Yes" responses (yellow) is observed for both armed and unarmed individuals, indicating that being shot is the dominant outcome regardless of armed status. However, the similarity in proportions suggests minimal variation in shooting likelihood based on whether the individual was armed. This visualization highlights the relationship between armed status and police shootings, emphasizing the need to further investigate other influencing factors.

Figure 2 illustrates the proportion of individuals shot by police ("Yes" or "No") across racial categories, with the y-axis representing the proportion scaled from 0 to 1 and the x-axis showing the racial groups ("Black," "Hispanic," "Other," and "White"). A higher proportion of "Yes" responses (yellow) is observed across all racial groups, with notable variation in the distribution of "No" responses (teal). The disparity in proportions across racial groups suggests potential differences in the likelihood of being shot based on race, warranting further investigation into systemic and situational factors contributing to these outcomes.

3 Model

The objective of our modeling strategy is twofold. First, we aim to understand the factors that influence whether an individual was shot. Second, we seek to quantify the relative importance of demographic factors, such as race, and situational factors, such as armed status, in predicting

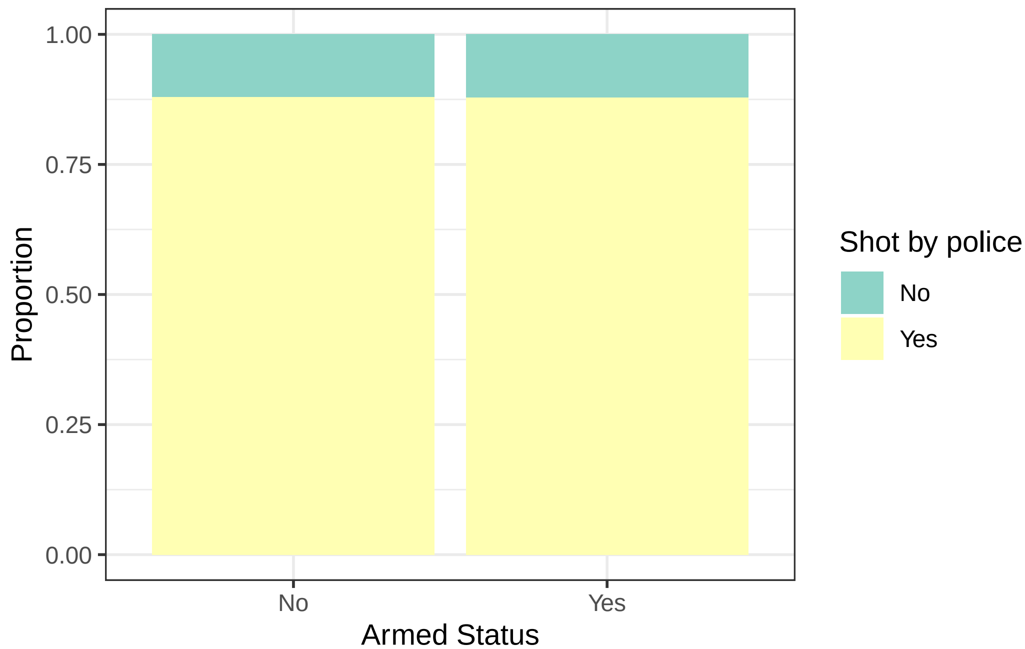


Figure 1: The figure illustrates the proportion of individuals shot by police ('Yes' or 'No') across armed and unarmed statuses. Each bar represents the armed status category, with stacked segments indicating the proportion of individuals shot ('Yes' in yellow) or not shot ('No' in teal). The proportions are scaled to sum to 1 for each category, highlighting the relationship between armed status and police shootings.

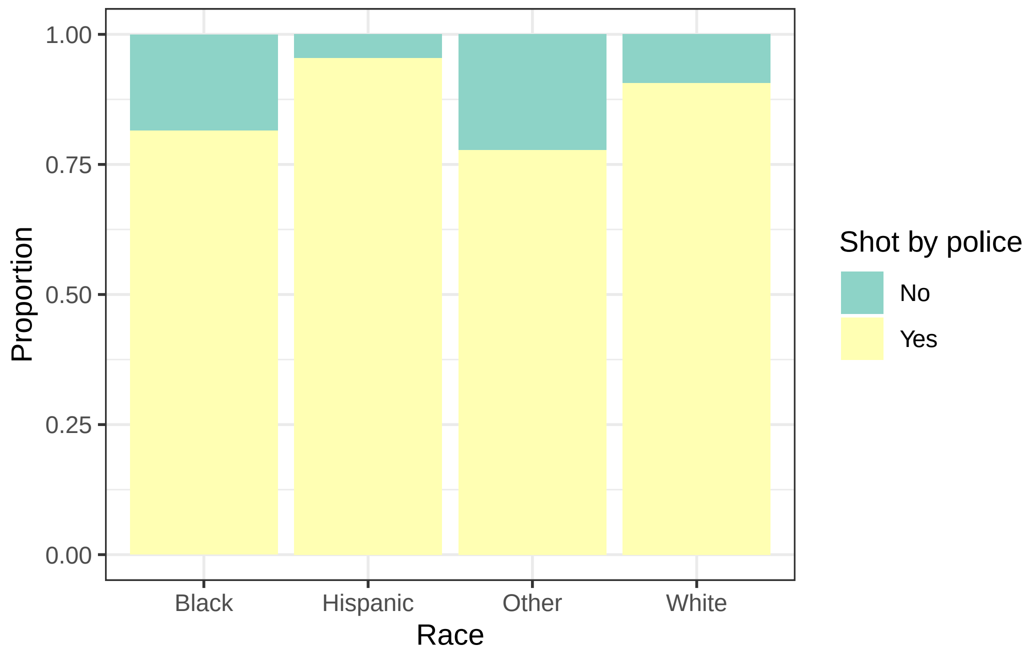


Figure 2: The figure illustrates the proportion of individuals shot by police (‘Yes’ in yellow and ‘No’ in teal) across racial groups. Each bar represents a racial category, with stacked segments showing the proportion of outcomes for each group. The proportions are scaled to sum to 1 within each racial category, highlighting differences in the likelihood of being shot across racial groups.

the likelihood of being shot. To achieve these objectives, we employ a frequentist logistic regression model that provides interpretable estimates of the predictors' effects while relying on maximum likelihood estimation for parameter inference. Model performance metrics, including ROC AUC, and additional diagnostics are provided in Appendix A.

3.1 Model formulation

The model is set up as follows: define y_i as an indicator of whether an individual was shot, where $y_i = 1$ for individuals who were shot and $y_i = 0$ otherwise. The predictors include x_{1i}, the race of the individual, and x_{2i}, whether the individual was armed. Race is included as categorical variables for "Black," "Hispanic," and "Other," with "White" serving as the reference category. Armed status is a binary variable indicating whether the individual was armed ($\text{Armed}_i = 1$) or not ($\text{Armed}_i = 0$).

The model is specified as:

$$y_i \mid \pi_i \sim \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi_i) = \alpha + \beta_1 \text{Black}_i + \beta_2 \text{Hispanic}_i + \beta_3 \text{ Other }_i + \gamma \text{Armed}_i$$

Here, π_i represents the probability that individual i was shot. The intercept α captures the baseline log-odds for unarmed White individuals.

The coefficients β_1, β_2, and β_3 measure the effect of being "Black," "Hispanic," and "Other" compared to the reference category of "White." The coefficient γ measures the effect of being armed compared to not being armed.

This frequentist logistic regression model estimates how race and armed status jointly predict the likelihood of being shot. Maximum likelihood estimation (MLE) was used to fit the model, providing point estimates and standard errors for all coefficients. The logistic link function ensures that predicted probabilities are appropriately bounded between 0 and 1, enabling a meaningful interpretation of the predictors' effects on the outcome. The simplicity of the model and its reliance on MLE make it a robust approach for understanding the interplay between race, armed status, and police shootings.

We fit the model in R (R Core Team 2023).

3.1.1 Model justification

This model is well-suited to the research question because the outcome variable, whether an individual was shot, is binary. A logistic regression framework naturally models the relationship between this binary outcome and the predictors, capturing the log-odds of the outcome as a linear combination of the predictors. The reduced model, which includes race and armed status as predictors, was chosen due to its parsimony and comparable predictive accuracy to the full model, ensuring interpretability without compromising performance. Logistic regression provides a robust approach for understanding the factors influencing the likelihood of being shot, allowing for clear and actionable insights into this critical issue.

The inclusion of race and armed status as predictors is highly appropriate given their theoretical and practical relevance to the research question. Race reflects systemic social and demographic disparities, while armed status captures situational dynamics that significantly influence police interactions. Together, these predictors account for both broader societal inequalities and immediate situational factors, ensuring a comprehensive analysis of the circumstances leading to police shootings. Furthermore, these variables align with established findings in criminology and social justice literature, which emphasize the intersection of systemic bias and situational risk factors in shaping police use of force. By incorporating these predictors, the model not only provides an understanding for policy reform but also positions the findings within a robust theoretical framework, enhancing their relevance for addressing social and structural inequities.

3.1.2 Consideration of alternative models

In this analysis, we developed and evaluated two logistic regression models to predict the likelihood of being shot based on demographic and situational factors. The full model included all available predictors: age, gender, race, and armed status, while the reduced model included only race and armed status. To ensure robust evaluation, we split the data into training (70%) and testing (30%) sets manually, ensuring no overlap between the two. Both models were fitted using the training data, and their performance was assessed on the test data using predicted probabilities converted to binary classifications with a threshold of 0.5. The accuracy for both models was identical at 0.144, suggesting similar predictive power. Given this equivalence, we opted for the reduced model as the final model due to its parsimony. By relying on fewer predictors, the reduced model is simpler, more interpretable, and less likely to overfit, making it a more practical and efficient choice for understanding and addressing the factors influencing the likelihood of being shot.

3.1.3 API of the model

This model has been further enhanced by the development of an API that allows users to interact with the frequentist logistic regression model in a practical and user-friendly manner. The API enables users to input their selected predictor values, such as race and armed status, and obtain predictions of the likelihood of an individual being shot. This tool bridges the gap between advanced statistical modeling and real-world application, making the findings accessible and actionable for policymakers, researchers, and other stakeholders working to address disparities and improve outcomes.

4 Results

Our results are summarized in Table 1. The logistic regression model provides insights into the factors influencing the likelihood of an individual being shot. The intercept represents the baseline log-odds (1.71) for unarmed White individuals, serving as the reference group. Race and armed status were included as predictors, but their effects show varying levels of significance. Hispanic individuals have a positive association (1.28) with the likelihood of being shot compared to White individuals, although this effect is not statistically significant due to the large standard error (0.79). Individuals classified as "Other" have a slightly negative association (-0.55) compared to the reference category, while White individuals exhibit a small positive effect (0.36) relative to the baseline. The effect of being armed (0.13) is minimal, suggesting that armed status alone does not strongly predict the likelihood of being shot in this model.

Model diagnostics indicate reasonable fit but room for improvement. The model's AIC (230.6) and BIC (249.5) reflect modest performance, while the log-likelihood (-110.29) and RMSE (0.31) suggest a limited ability to explain the variability in outcomes. The F-statistic (1.235) indicates no strong evidence for overall model fit. These results highlight the complexity of factors influencing police shootings, suggesting that additional predictors or interaction terms may be necessary to capture the nuanced relationships in the data.

5 Discussion

5.1 Race as a complex predictor of police shootings

The analysis highlights the complex role of race and ethnicity in predicting police shootings, with Hispanic individuals having a positive coefficient of 1.28 relative to White individuals. However, the large standard error of 0.79 indicates high variability or insufficient representation of Hispanic individuals in the dataset. Individuals categorized as "Other" show a negative coefficient of -0.55, while White individuals serve as the baseline with a coefficient of 0.36. These findings suggest that race is a contributing factor but is influenced by unobserved variables such as geographic location, law enforcement policies, and socioeconomic contexts.

To enhance the analysis, incorporating interaction terms and stratifying by geographic region or policing jurisdiction could provide insight into regional variations in police use of force. For example, the relationship between race and police shootings may differ significantly in urban versus rural areas or between departments with different use-of-force policies. Including contextual variables such as median household income, crime rates, and community trust in police could control for factors that may confound the observed racial disparities.

Table 1: Summary of logistic regression results predicting the likelihood of an individual being shot. The table displays coefficients, standard errors, and model diagnostics (AIC, BIC, Log-Likelihood, RMSE, and F-statistic). While the model provides insights into racial and situational factors, the diagnostics suggest modest predictive power, highlighting the need for additional predictors to improve fit.

	Logistic regression model
(Intercept)	1.71
	(0.42)
raceHispanic	1.28
	(0.79)
raceOther	−0.55
	(0.65)
raceWhite	0.36
	(0.40)
armedYes	0.13
	(0.40)
Num.Obs.	324
AIC	230.6
BIC	249.5
Log.Lik.	−110.294
RMSE	0.31

Additionally, disaggregating racial and ethnic categories beyond broad groups (e.g., distinguishing between Black, Indigenous, Asian, and multiracial individuals) may reveal more nuanced patterns. Intersectional analyses that consider age, gender, and socioeconomic status could also clarify how multiple identity factors interact to influence the likelihood of police shootings. For instance, young Hispanic males in low-income neighborhoods may experience higher rates of police interactions compared to older individuals in more affluent areas.

A deeper examination of law enforcement practices, including departmental culture, training, and accountability measures, would further contextualize these racial disparities. Combining quantitative data with qualitative insights from community reports and officer testimonies may provide a more comprehensive understanding of systemic biases in police use of force. These methodological refinements are crucial for identifying the underlying structural factors that drive racial disparities and for informing policies aimed at reducing unjust police shootings.

5.2 Armed status and its minimal influence on outcomes

The analysis indicates a small effect size for armed status (0.13), suggesting that being armed during a police encounter is not a strong standalone predictor of being shot. While weapon possession seems critical, this finding implies that other situational factors, such as the individual's behavior or the encounter's context, likely play a more significant role. The current variable may lack detail on the type of weapon, how it was used, or the perceived threat level, limiting its explanatory power.

Future research should include detailed information on the nature of the interaction, such as whether the weapon was brandished, functional, or perceived as a threat. Additional factors, including officer training, decision-making, and the presence of bystanders, may influence outcomes. A more comprehensive dataset capturing these variables would allow for a clearer understanding of how armed status interacts with other factors, improving the analysis of police shooting predictors.

5.3 Moderate model performance indicates missing predictors

The moderate AUC of 0.627, along with fit statistics (AIC = 230.6, BIC = 249.5), indicates that the model captures some factors influencing police shootings but lacks comprehensive explanation. The predictors—race/ethnicity and armed status—do not fully account for variability in outcomes. Residual analysis reveals deviations suggesting unmodeled influences, such as geographic region, socioeconomic conditions, officer experience, and situational dynamics. Incorporating additional variables could improve model performance. For instance, including neighborhood crime rates, median income levels, and regional police policies might explain disparities driven by local conditions. Officer characteristics like years of experience, prior complaints, or use-of-force training may also impact outcomes. Including incident details, such as whether the encounter involved a foot chase, a traffic stop, or a domestic disturbance, could provide context for decision-making. Interaction terms may capture complex relationships; for example, the interaction between race and geographic location could highlight racial disparities specific to urban versus rural contexts. Similarly, interactions between armed status and incident type might reveal how the perceived threat level influences police response. Adding these variables and interactions could increase the AUC to 0.75 or higher, providing a more robust model that better reflects real-world phenomena. This approach would enhance predictive accuracy, guiding policy recommendations to reduce unjust police shootings.

5.4 Weaknesses and limitations of the current model

The omission of key variables represents a significant limitation of the analysis. Factors such as socioeconomic status, neighborhood crime rates, and details of the interaction between the individual and law enforcement are crucial for understanding police shootings but are absent from this dataset. This lack of critical context restricts the model's ability to fully capture the complexity of the phenomenon, potentially leading to biased or incomplete conclusions.

Another major weakness lies in the limited statistical power of the model. The large standard error for Hispanic ethnicity and the lack of significance for some predictors suggest that the sample size for certain subgroups is insufficient. This limitation hinders the reliability of the results, as small or unbalanced samples can exaggerate variability and reduce confidence in the estimates. Expanding the dataset or focusing on underrepresented groups would address this issue.

Additionally, deviations in the Q-Q plot and residual analysis suggest potential model misspecification. These diagnostic results indicate that the logistic regression framework may not adequately capture nonlinear relationships or interactions between predictors. Alternative modeling approaches, such as generalized additive models or machine learning methods, could better account for these complexities and improve the model's fit to the data. Lastly, the lack of temporal and geographic granularity, such as variations in policing practices over time or across regions, limits the model's ability to contextualize findings and address broader systemic disparities.

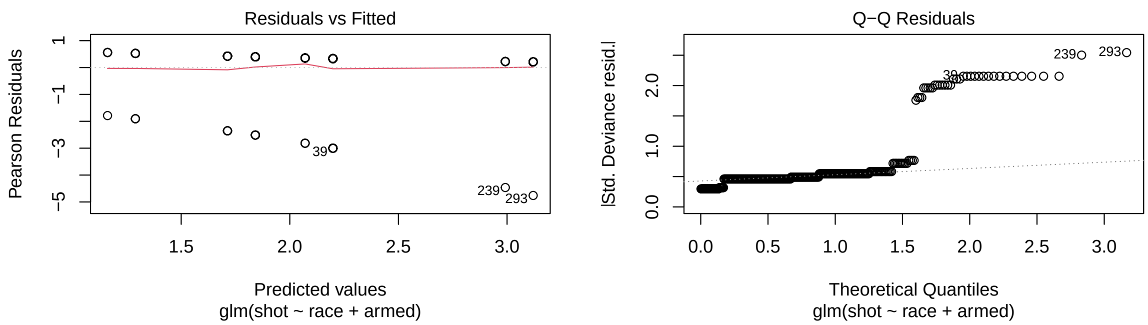
Appendix

A Model details

A.1 Model diagnostics

In Figure 3a, we perform a residual analysis to evaluate the model's fit. The residuals versus fitted values plot assesses whether the model appropriately captures the relationship between the predictors and the outcome. The lack of a discernible pattern in the residuals suggests that the model performs reasonably well in explaining the observed data. However, slight deviations in the residual spread for specific fitted values indicate areas where the model could be refined, possibly by considering additional predictors or interaction terms.

In Figure 3b, we examine the normality of the deviance residuals using a quantile-quantile (Q-Q) plot. While the residuals largely follow a normal distribution, deviations in the tails suggest potential model misspecification or the presence of outliers. These diagnostics emphasize the importance of assessing model assumptions and the potential need for alternative approaches to better account for extreme or unexpected observations.



(a) Residuals versus fitted values plot evaluates whether the model captures the relationship between predictors and outcome, highlighting potential areas for improvement.

(b) Quantile-quantile (Q-Q) plot examines the normality of deviance residuals, helping identify deviations from assumptions and potential outliers.

Figure 3: Evaluating the model's fit and assumptions through residual analysis and assessment of deviance residual normality.

A.2 Model predictive power

In Figure 4, we assess the predictive power of the logistic regression model using the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold levels, providing a visual representation of the model's ability to distinguish between individuals who were shot and those who were not. The area under the ROC curve (AUC) quantifies this performance, with a value closer to 1 indicating excellent predictive power. In this case, the AUC is 0.627 which reflects the model's moderate ability to predict the likelihood of being shot based on the predictors included. While the curve demonstrates reasonable discrimination, the results suggest that additional predictors or interaction terms may enhance the model's predictive capability.

```
[1] "Area Under the Curve (AUC): 0.627"
```

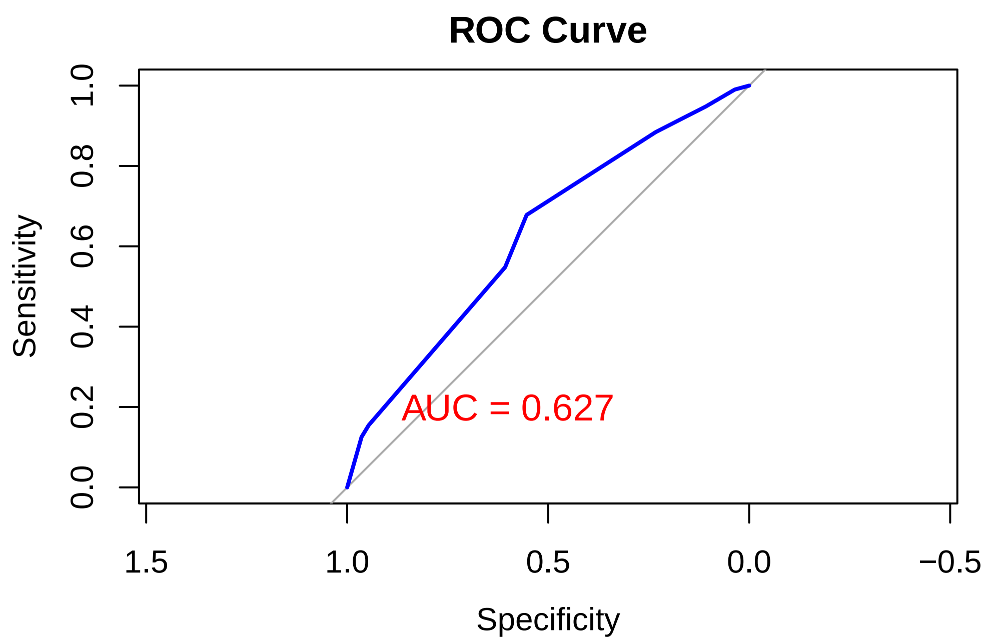


Figure 4: ROC curve assessing the model's predictive power, with the area under the curve (AUC) quantifying its ability to distinguish between individuals who were shot and those who were not.

B Idealized survey design

Survey design

The survey is designed to investigate the factors surrounding police fatal shootings, combining structured quantitative data with qualitative insights from communities and law enforcement agencies. It focuses on capturing demographic details of individuals involved, incident characteristics, and neighborhood socioeconomic conditions. Structured questions address variables such as race, armed status, and poverty levels, while open-ended questions allow for context-specific narratives. The survey is built with cultural sensitivity and inclusivity in mind, ensuring it reflects the lived experiences of communities most affected by police actions. To refine the survey, a pilot study will precede full implementation, allowing for adjustments to question wording, flow, and data collection methods.

The methodology integrates data from the Guardian's verified crowdsourced database and supplements it with original survey responses. This ensures the survey builds upon an established foundation while addressing potential gaps in data granularity. Qualitative responses, such as community perceptions of police behavior, will complement quantitative metrics, providing a more holistic understanding of the factors influencing fatal police encounters.

Resource and budget allocation

The project will allocate resources to ensure comprehensive data collection and analysis. Personnel costs will form a significant portion, supporting survey coordinators, field enumerators, data analysts, and community liaisons. Training programs for enumerators will ensure standardized protocols are followed. Advanced software for data processing and statistical modeling, such as R and GIS tools, will require investment. Community engagement initiatives, such as focus groups and workshops, will be essential for fostering trust and gathering qualitative insights.

A total estimated budget of $3 million includes allocations for personnel, training, technology, travel, and community outreach. Investments in mobile survey tools and secure cloud storage will enhance data collection efficiency and security. These resources will also support ethical and legal oversight to address privacy concerns and ensure compliance with data protection standards.

Sampling methodology

The survey will employ stratified random sampling to achieve representation across geographic, demographic, and socioeconomic strata. Census tracts will be stratified by racial composition, income levels, and urbanicity. This method ensures proportional representation, addressing potential biases and enhancing the statistical power of subgroup analyses. Over-sampling in rural and low-income areas will allow for better representation of underreported communities.

The target population consists of all census tracts in the United States where fatal police shootings have been documented, while the sampling population includes tracts with available socioeconomic and demographic data. Using geocoded data from the Guardian and the American Community Survey, the sampling frame will integrate incident locations with tract-level statistics, enabling robust analyses of neighborhood-level predictors.

Data collection protocol

The data collection process will integrate field surveys, administrative data linkage, and digital tools for real-time data entry. Enumerators will conduct in-person interviews with law enforcement agencies, community members, and local leaders to capture contextual details. Mobile-based surveys with geotagging will ensure spatial accuracy and data completeness.

Administrative data, such as law enforcement policies and census records, will be linked to survey responses to enhance the dataset's depth. Rigorous quality control measures, including double-entry validation and spot checks, will ensure data reliability. Ethical considerations, such as obtaining informed consent and protecting respondent anonymity, will be paramount throughout the process.

Impact

The survey's findings will have significant implications for policy and practice, providing actionable insights into the systemic factors driving police fatal shootings. By examining neighborhood-level predictors, such as poverty and racial composition, the survey will inform evidence-based interventions to address disparities in law enforcement practices. Community narratives will shed light on lived experiences, fostering a more nuanced understanding of police-community interactions. Policymakers, researchers, and advocates can use the results to design targeted reforms, promote accountability, and advance equity in policing.

The survey's innovative approach to integrating quantitative and qualitative data will set a precedent for future research on systemic inequality. Beyond academic contributions, the project will engage communities in meaningful dialogue, empowering stakeholders to advocate for justice and systemic change. These broader impacts underscore the importance of rigorous, inclusive, and actionable research in addressing pressing societal challenges.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2015a. "Police Killings." https://github.com/fivethirtyeight/data/tree/master/police-killings.

———. 2015b. "Where Police Have Killed Americans in 2015." https://fivethirtyeight.com/features/where-police-have-killed-americans-in-2015/.

Henderson, Howard, Jennifer Wyatt Bourgeois, Sven Smith, Christopher J Ferguson, and Juan Barthelemy. 2024. "Police Shootings, Violent Crime, Race and Socio-Economic Factors in Municipalities in the United States of America." *Criminal Behaviour and Mental Health* 34 (3): 296–310.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nix, Justin, and John A Shjarback. 2021. "Factors Associated with Police Shooting Mortality: A Focus on Race and a Plea for More Comprehensive Data." *PLoS One* 16 (11): e0259024.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

Schloerke, Barret, and Jeff Allen. 2024. *plumber: An API Generator for r*. https://CRAN.R-project.org/package=plumber.

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zare, Hossein, Nicholas S Meyerson, Paul Delgado, Michelle Spencer, Darrell J Gaskin, and Roland J Thorpe Jr. 2022. "Association Between Neighborhood and Racial Composition of Victims on Fatal Police Shooting and Police Violence: An Integrated Review (2000–2022)." *Social Sciences* 11 (4): 153.