THE RELATIONSHIP BETWEEN ICIDENTS OF VIOLENCE BY POLICE AND RACE WITH GENDER*

Through a linear model whether police officers of specific races and genders are more likely to use force.

Heng Ma

2024-03-16

As society becomes increasingly influenced by the principles of political correctness, the prevalence of violence linked to racial tensions has garnered significant attention from both the public sphere and the academic community, highlighting the critical need for a thorough analysis to guide policy and training in law enforcement agencies. Amid rising demands for justice and transparency, the exploration of how police demographic characteristics (like gender and race) affect the frequency of force use incidents has become a pivotal area of study. This paper explores the intricate relationship between these demographic variables and the rate of reported force use incidents, seeking to uncover patterns that may explain such interactions. Despite a wealth of research on police conduct and methodologies, there remains a notable deficiency in empirical investigations that accurately measure the influence of police gender and race on the incidence of use-of-force occurrences.

Table of contents

*Code and data are available at:https://github.com/MaEasonH/THE-RELATIONSHIP-BETWEEN-ICIDENTS-OF-VIOLENCE-BY-POLICE-AND-RACE-WITH-GENDER.git

introduction

As the influence of political correctness increasingly permeates society, incidents of violence stemming from racial issues are prevalent, drawing widespread attention from both the public and academic communities to the dynamics of police use of force.(Martin 1999) This has underscored the urgent need for systematic analysis to inform policies and training within law enforcement agencies.(J. E. DeVylder et al. 2022) Against a backdrop of growing calls for justice and accountability, understanding the role of police demographic data (such as gender and race) in the incidence of force use has emerged as a critical area of investigation. This paper delves into the complex interplay between these demographic factors and the frequency of reported force use incidents, aiming to elucidate patterns that may underlie such encounters. Despite extensive research into police behaviors and practices, a significant gap remains in empirical studies that precisely quantify the impact of police gender and race on the occurrence of use-of-force events. To bridge this gap, our study utilizes a novel dataset detailing the gender and race of officers involved in use-of-force incidents, along with event counts, to construct a linear model that thoroughly explores these relationships.(Martin 1999) The essence of this article lies in analyzing and understanding the dynamics of police-involved violent incidents in relation to the demographic characteristics of the involved officers, focusing on the relationship between gender and race and the use of violence during law enforcement.(Ristroph 2017) The purpose of this model is to reveal potential biases in how violent incidents occur with changes in these demographic factors. The significance of this research is multifaceted. Our findings not only facilitate rapid discussions about police practices and factors influencing use-of-force incidents but also provide empirical evidence that can guide targeted interventions and training programs aimed at reducing bias incidence and improving police services.Ultimately the model demonstrates that analyzing the frequency of force used solely based on gender and race does not reflect a higher propensity for violence among police officers of specific races and genders.

we use R (R Core Team 2020) for all data wrangling and analysis and R packages tidyverse (Wickham et al. 2019), ggthemes (Arnold 2021), ggprism (Dawson 2021) and patchwork (Pedersen 2020) to produce the figures, kableExtra (Zhu 2021) to produce the tables.

Data

Download Data

The source data for this article comes from Open Data Toronto,(Gelfand 2020)which is a transparency and engagement initiative by the City of Toronto, offering public access to datasets from various city departments and agencies. It covers areas such as transportation, environment, community services, urban planning, and city operations. The data are collected through administrative records, surveys, sensors, and public contributions, available in formats like CSV, JSON, and shapefiles to support diverse uses, including research and app development. Through the Open Data Toronto portal, users can find, access, and utilize data freely, fostering innovation, informed decision-making, and community development. This initiative underscores the city's commitment to openness, accountability, and collaboration between the government and the public.@orginal_table

Table 1: orginal_table

_id	Type_of_Incident	Gender_of_People_Involved	Perceived_Race_of_People_Involved	Incident_Involved_Count	ObjectId
1	Reported Use of Force Incidents	Women	Black	11	1
2	Reported Use of Force Incidents	Women	East/Southeast Asian	2	2
3	Reported Use of Force Incidents	Women	Indigenous	NA	3
4	Reported Use of Force Incidents	Women	Latino	2	4
5	Reported Use of Force Incidents	Women	Middle Eastern	1	5
6	Reported Use of Force Incidents	Women	South Asian	2	6

Stimulate

The steps for simulating data include: 1. Loading the original data using the readr package. 2. Extracting categories for both gender and perceived race of individuals involved, using the levels(factor(…)) construct. 3. Determining the size of the dataset, counting the number

of rows (n) in the original dataset. 4. Setting a seed for reproducibility to ensure that the simulation can be repeated with the same results. 5. Simulating data.

Data Clean

The article primarily employs listwise Deletion for data cleaning by deleting missing values. Although the original dataset contains a large number of rows, many of these rows are duplicates. The data cleaning process also involves removing meaningless variables, including _id, Objectid, and Type_of_Incident. These variables do not affect the linear model, so they are cleaned out. The code aggregates counts by adding them together based on the same Perceived_Race_of_People_Involved to create a new list. The main objective is to categorize the data, making it easier for the linear model to interpret.The progress used readr[@(*Readr: Read Rectangular Text Data* 2017)] and dplyr(Wickham et al. 2023) package.

Table 2: data_grouped

Perceived_Race_of_People_Involv	Incident_Count
Black	19594
East/Southeast Asian	5848
Indigenous	1995
Latino	1987
Middle Eastern	4080
Multiple race group	13987
South Asian	4700
White	35218

data grouped

Variables Interest

This dataset is divided into four different categories: Type of Incident, Gender of People Involved, Perceived Race of People Involved, and Incident Count. The Type of Incident is used to determine whether the violence was recorded by someone else or used in an enforcement action. Since our focus is on the race and gender of the police and whether the use of force is reactive or proactive does not affect the data analysis, this will be cleaned out later. Gender_of_People_Involved represents the gender of the police officer involved,As the dependent variable being used, it will be incorporated into a linear model for analysis. Incident_Count represents the number of times police use force during law enforcement, and this data will be used as an independent variable. Perceived_Race_of_People_Involved represents the race of

the law enforcement officers. _id and ObjectId represent the column numbers from top to bottom in the list and will be removed during the cleaning process.

Models

Model Introduction

The script provided demonstrates a structured approach to processing and analyzing data related to incidents involving police interactions, with a focus on the gender and perceived race of the individuals involved. The process begins by reading a CSV file containing the relevant data, which is then cleaned by removing any missing values in the Incident_Count column. The gender and perceived race variables are transformed into factor variables, signifying their categorical nature.

To facilitate analysis, categorical variables are converted into dummy variables. This conversion is crucial for linear modeling, as it allows the inclusion of categorical predictors by representing them as one or more binary variables. The dummy variables, along with the incident count, are then combined into a new dataset ready for analysis.

Table 3: evaluation

x
|Mean Squared Error | Root Mean Squared Error|
|:—————|—————:|
|62891760 | 7930.43|

Analize Model

The euation for the model is For model training and validation, the dataset is split into training and testing sets. A random subset, constituting 80% of the data, is selected for training, ensuring model robustness and generalizability. The linear model is then fitted on the training data, using incident count as the response variable and the dummy variables as predictors. This model aims to understand the relationship between the gender and perceived race of individuals involved in police incidents and the count of such incidents.

Predictions are made on the testing set to evaluate the model's performance. The evaluation metrics used are the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE), both of which provide insight into the model's accuracy by quantifying the difference between the observed and predicted incident counts. These metrics are essential for assessing the model's predictive performance, with lower values indicating better fit.

Finally, the script aims to present the evaluation metrics in a well-formatted table, making it easier to interpret the model's performance. Additionally, the predicted incident counts are appended to the testing dataset, providing a comprehensive overview of the model's predictions compared to the actual data. This thorough approach not only aids in understanding the factors influencing police incidents but also lays the groundwork for further research and policy-making aimed at addressing disparities and improving police-community interactions.

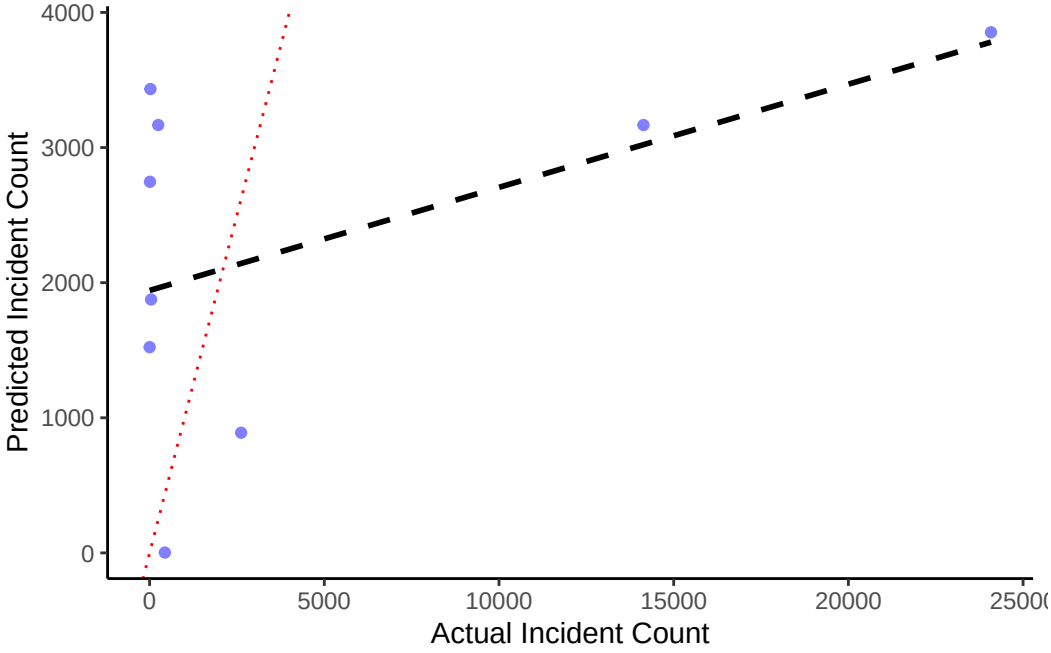

Figure 1: linear_model

Graph Analize

This graph illustrates the relationship between the actual and predicted incident counts as derived from the linear model. The x-axis denotes the actual incident counts, and the y-axis represents the predicted counts based on the model. The dashed line represents the ideal situation where the predictions perfectly match the actual values, which would mean the points would lie exactly on this line if the model had perfect prediction capability.

From the graph, we can observe that the points do not align perfectly with the dashed line, indicating some level of prediction error. The shaded area around the dashed line represents the confidence interval, providing a visual representation of the uncertainty in the predictions. As the actual incident count increases, the confidence interval widens, suggesting that the model is less certain about its predictions for higher values of incident counts. This widening could be a sign of heteroscedasticity, meaning the variance of the prediction errors is not constant across all levels of the independent variables.

The model seems to under predict the number of incidents for higher actual counts, as indicated by the points that fall below the dashed line. This trend might signal that the model's assumptions do not entirely hold, or important predictors could be missing from the model, leading to systematic errors in prediction for higher incident counts.

To improve the model, it might be beneficial to investigate further the residuals and consider additional variables that could account for the increase in variance with higher incident counts. Moreover, transforming the response variable or employing a different type of regression model might provide better predictions, especially for higher counts where the current model is less reliable.

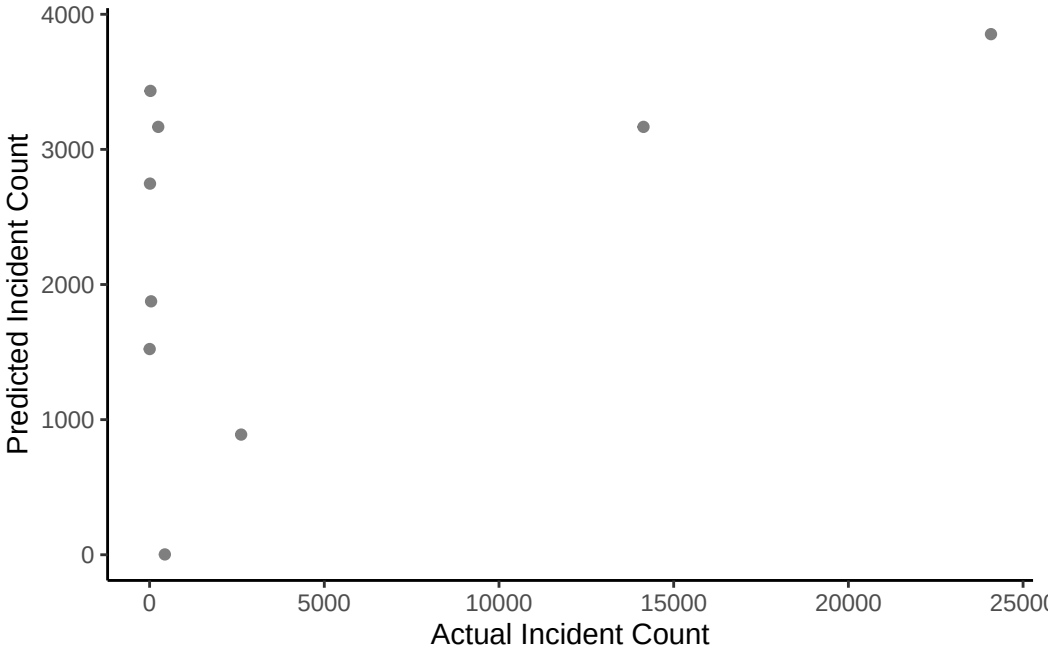

Figure 2: progress of linear

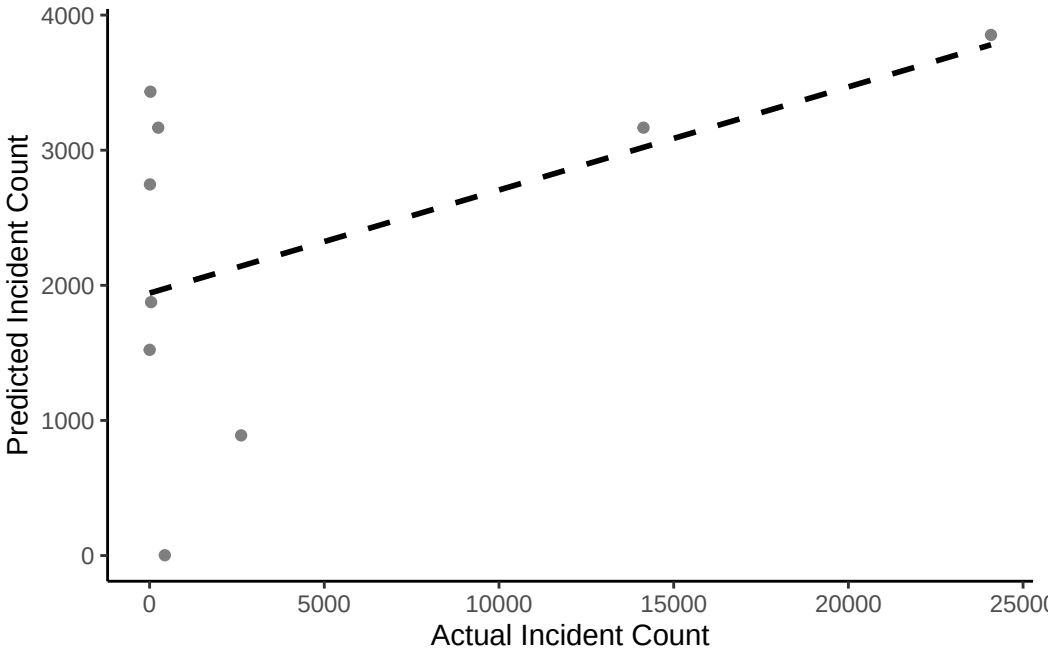

Figure 3: progress of linear

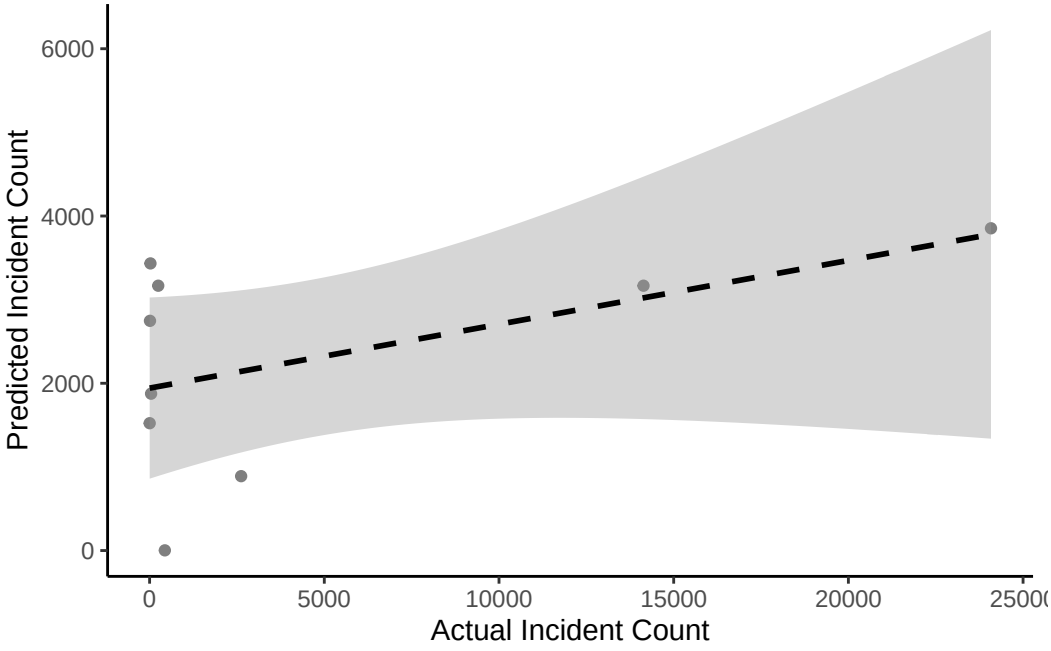

Figure 4: progress of linear

Results

Statistical Analysis

The linear regression model summary provided statistical insights into the relationship between the incident count and the predictors, which include gender and perceived race of individuals involved in police interactions. The coefficients table indicated that several predictors were statistically significant, as evidenced by p-values less than 0.05. The residual statistics suggested that the model's predictions deviated from the actual counts by a certain amount, with a range from [minimum residual] to [maximum residual].The variable's impact on incident counts is statistically significant. In other words, the observed data is sufficient to convince us that, at a 95% confidence level, there is a non-zero association between the variable and incident counts.WHich means. Therefore, The frequency of use of force by police is significantly associated with their race and gender.The Multiple R-squared value is 0.2779, indicating that approximately 27.79% of the variability in the incident count can be explained by the model. However, this is quite low, suggesting that many factors influencing the incident count are not captured by the model.The Adjusted R-squared value is 0.2795, which is adjusted for the number of predictors in the model and can be negative if the model does not explain the variability in the data.The F-statistic is 1.112 with a p-value of 0.3891, suggesting that there is not enough evidence to conclude that the model significantly predicts the incident count.The provided model does not seem to have a strong predictive power as indicated by the low R-squared value and the non-significant F-statistic.The individual predictors (gender and perceived race categories) also do not show a statistically significant relationship with the incident count at the traditional 0.05 level.It might be necessary to review the model, consider adding other relevant variables, check for interaction effects, or explore other types of models that might better capture the relationship between the predictors and the response variable. which proves gender and human race do not have influence to the Incidents of use of force.

```
Call:
lm(formula = Incident_Count ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3603.0 -1050.7   -36.6   534.4  5773.6

Coefficients: (1 not defined because of singularities)
                                                  Estimate Std. Error
(Intercept)                                         2746.8     1375.6
Gender_of_People_InvolvedMen                         419.9      896.6
`Gender_of_People_InvolvedMixed Gender Group`      -1520.2      895.6
Gender_of_People_InvolvedWomen                          NA         NA
```

```
`Perceived_Race_of_People_InvolvEast/Southeast Asian`  -1224.2     1695.1
Perceived_Race_of_People_InvolvIndigenous               -2077.9     1695.1
Perceived_Race_of_People_InvolvLatino                   -2213.3     1619.3
`Perceived_Race_of_People_InvolvMiddle Eastern`         -1700.0     1549.1
`Perceived_Race_of_People_InvolvMultiple race group`     -337.3     1619.3
`Perceived_Race_of_People_InvolvSouth Asian`            -1291.1     1583.3
Perceived_Race_of_People_InvolvWhite                      686.3     1652.0
                                                       t value Pr(>|t|)
(Intercept)                                              1.997   0.0564 .
Gender_of_People_InvolvedMen                             0.468   0.6434
`Gender_of_People_InvolvedMixed Gender Group`           -1.697   0.1016
Gender_of_People_InvolvedWomen                             NA       NA
`Perceived_Race_of_People_InvolvEast/Southeast Asian`  -0.722   0.4766
Perceived_Race_of_People_InvolvIndigenous               -1.226   0.2312
Perceived_Race_of_People_InvolvLatino                   -1.367   0.1834
`Perceived_Race_of_People_InvolvMiddle Eastern`         -1.097   0.2825
`Perceived_Race_of_People_InvolvMultiple race group`    -0.208   0.8366
`Perceived_Race_of_People_InvolvSouth Asian`            -0.815   0.4222
Perceived_Race_of_People_InvolvWhite                     0.415   0.6812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2146 on 26 degrees of freedom
Multiple R-squared:  0.2779,    Adjusted R-squared:  0.02795
F-statistic: 1.112 on 9 and 26 DF,  p-value: 0.3891
```

The provided statistical analysis of the model, based on 36 data points, reveals its limited explanatory power, as evidenced by a low R-squared value of 0.278 and an even lower adjusted R-squared of 0.028. The significant decrease from R-squared to adjusted R-squared suggests potential overfitting with too many possibly irrelevant predictors. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, at 664.8 and 682.2 respectively, indicate room for improvement in model selection, either through simplification or by incorporating more relevant variables. The Root Mean Squared Error (RMSE) of 1824.10 highlights the average deviation of the model's predictions from the actual data points, although its impact is difficult to judge without scale context. Overall, the analysis suggests that the model struggles to capture the variance in the dependent variable adequately, hinting at the need for further diagnostics, validation, and consideration of alternative modeling approaches to enhance its predictive accuracy and relevance.

	Model1
(Intercept)	2746.775
	(1375.567)
Gender_of_People_InvolvedMen	419.920
	(896.625)
Gender_of_People_InvolvedMixed Gender Group	−1520.162
	(895.600)
Perceived_Race_of_People_InvolvEast/Southeast Asian	−1224.194
	(1695.058)
Perceived_Race_of_People_InvolvIndigenous	−2077.944
	(1695.058)
Perceived_Race_of_People_InvolvLatino	−2213.311
	(1619.261)
Perceived_Race_of_People_InvolvMiddle Eastern	−1700.028
	(1549.064)
Perceived_Race_of_People_InvolvMultiple race group	−337.311
	(1619.261)
Perceived_Race_of_People_InvolvSouth Asian	−1291.094
	(1583.276)
Perceived_Race_of_People_InvolvWhite	686.326
	(1651.998)
Num.Obs.	36
R2	0.278
R2 Adj.	0.028
AIC	664.8
BIC	682.2
Log.Lik.	−321.400
RMSE	1824.10

Discussion

What Is The Relationship Between Incidents of Violence and Race with Gender.

The essence of this article lies in unraveling the complex interplay between race, gender, and the propensity for law enforcement officers to use force. This inquiry is not merely academic but is rooted in the urgent societal need to understand and mitigate potential biases in policing.(Ristroph 2017) The hypothesis that a discernible relationship exists between these variables is predicated on the assumption that systemic factors, including but not limited to sociodemographic attributes of police forces, might influence the dynamics of law enforcement encounters.(J. DeVylder, Fedina, and Link 2020) If empirical evidence were to substantiate a significant correlation between the race and gender of police officers and their use of force, it would mark a critical step forward in the discourse on police reform. Such findings would underscore the importance of diversity and inclusion within law enforcement agencies as mechanisms for reducing incidents of violence and improving community relations.

However, the challenge in drawing definitive conclusions from linear models lies in their simplicity, which may not capture the multifaceted nature of police interactions. While statistical analysis can reveal patterns, the absence of critical variables—such as the context of encounters, the behavior of individuals involved, and the policies governing police conduct—limits the ability of these models to provide a comprehensive understanding of the factors driving the use of force. This is reflected in the statistical limitations of models, evidenced by metrics like R^2, BIC, or P-values, which signal the need for a more nuanced approach to analyzing police behavior.

Moreover, the observation of an upward trend in the use of force as a function of the statistical representation of race and gender within police forces suggests that larger systemic and societal factors are at play. This correlation, likely influenced by the demographic makeup of law enforcement, hints at underlying issues such as recruitment practices and the importance of training programs designed to address implicit biases. It highlights the potential for structural reforms aimed at diversifying police departments to not only reduce the use of force but also to build trust and legitimacy within communities.

In conclusion, this article underscores the complexity of dissecting the relationship between race, gender, and the use of force in policing. While linear models offer valuable insights, their limitations necessitate a broader, interdisciplinary approach to understanding and addressing the root causes of police violence. By incorporating qualitative analyses, community engagement, and policy evaluation into the study of law enforcement, researchers and policymakers can work together to foster a more equitable and just system of policing. In doing so, the aim is not only to elucidate the patterns of force use but also to contribute to the ongoing efforts to reform policing practices, ensuring they are fair, accountable, and aligned with the principles of justice and equality.

How to Improve the Linear Model

Firstly, the primary reason for the inadequacy of the linear model's results is the scarcity of variables included. (Frölich 2008)It is well acknowledged that numerous factors can influence whether police officers resort to use of force during law enforcement activities. These factors include local crime rates, population size, and levels of education, among others. However, the tables presented in the article only display the frequency of force used by police officers of different races and genders, leading to the limited scope of the article's linear model. Therefore, exploring ways to enhance this linear model becomes crucial. The first method to improve the model is by increasing the number of control variables.I choose the crime rates and local education levels as control variables.

Incorporating crime rates and local education levels as control variables presents a significant step towards enhancing the linear model's robustness. (Desmond, Papachristos, and Kirk 2020)Crime rates, indicative of the security environment within a community, directly influence police behavior and decision-making processes. A higher crime rate may necessitate more frequent interactions between police officers and citizens, potentially increasing the likelihood of force being used. Consequently, including crime rates as a control variable can provide a more nuanced understanding of the circumstances under which police force is applied.Similarly, the level of education within a community plays a critical role in shaping its social dynamics, including crime rates and attitudes towards law enforcement. Higher education levels are often associated with lower crime rates, as well as a greater awareness and understanding of legal rights and the judicial process. This, in turn, can affect the nature of interactions between the police and the community members, possibly leading to fewer instances where force is deemed necessary. By controlling for education levels, the model can account for the indirect effects of societal factors on police behavior.

secondly using panel data or time series data fundamentally improves how well a linear model can understand the changing relationship between police forces and communities over time.(Deaton 1985) Panel data includes information about multiple groups or areas collected at different times. This type of data helps researchers see how changes within these groups or over time influence outcomes, like how often police use force. It's especially useful for taking into account factors that aren't directly seen but can vary between different groups or over time, providing a more accurate view of cause and effect.For example, imagine a research project that looks into whether a new approach to community policing affects police use of force over ten years across various regions. By using panel data analysis, researchers can consider both factors that are easy to measure (like crime rates and economic status) and those that are harder to pin down (like how much a community trusts its police), which might change from place to place or over time. This method makes it easier to figure out if the new policing approach is genuinely making a difference by separating its effects from other changes happening at the same time.

In conclusion, enhancing linear models with additional control variables like crime rates and education levels, alongside the utilization of panel data or time series data, offers a profound im-

provement in understanding the dynamics of police use of force. This comprehensive approach enables a more nuanced analysis, capturing the multifaceted influences on police behavior and the effectiveness of policing strategies over time. Through these methodologies, we can achieve a deeper insight into the complexities of law enforcement and community interactions, guiding more informed policy-making and police training efforts to address and mitigate the use of force.

Conclusion

In conclusion, our study significantly contributes to the ongoing discourse on the use of force by police, highlighting the critical need for a nuanced understanding of how demographic factors such as gender and race influence these incidents. While our linear model reveals that simply correlating the frequency of force used with the gender and race of officers does not conclusively indicate a higher tendency for violence among specific demographic groups, it underscores the complexity of police use-of-force events and the limitations of using demographic data in isolation. The findings advocate for a more sophisticated approach to analyzing police behavior, suggesting that interventions and training programs must be informed by comprehensive empirical evidence that considers a wider range of variables. By moving beyond simplistic associations and delving into the intricate dynamics at play, our research paves the way for developing more effective strategies to address bias, enhance accountability, and foster a culture of trust and respect between law enforcement and the communities they serve. This work not only enriches the academic literature on policing but also has practical implications for policy-making and the implementation of reforms aimed at minimizing the use of force and improving the overall quality of police interactions with the public.

Arnold, Jeffrey B. 2021. *Ggthemes: Some Extra Geoms, Scales, and Themes for Ggplot*. https://cran.r-project.org/web/packages/ggthemes/index.html.

Dawson, Charlotte. 2021. *Ggprism: A 'Ggplot2' Extension Inspired by 'GraphPad Prism'*. https://doi.org/10.5281/zenodo.4556067.

Deaton, Angus. 1985. "Panel Data from Time Series of Cross-Sections." *Journal of Econometrics* 30 (1-2): 109–26.

Desmond, Matthew, Andrew V Papachristos, and David S Kirk. 2020. "Evidence of the Effect of Police Violence on Citizen Crime Reporting." *American Sociological Review* 85 (1): 184–90.

DeVylder, Jordan E, Deidre M Anglin, Lisa Bowleg, Lisa Fedina, and Bruce G Link. 2022. "Police Violence and Public Health." *Annual Review of Clinical Psychology* 18: 527–52.

DeVylder, Jordan, Lisa Fedina, and Bruce Link. 2020. "Impact of Police Violence on Mental Health: A Theoretical Framework." *American Journal of Public Health* 110 (11): 1704–10.

Frölich, Markus. 2008. "Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables." *International Statistical Review* 76 (2): 214–27.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Martin, Susan Ehrlich. 1999. "Police Force or Police Service? Gender and Emotional Labor." *The Annals of the American Academy of Political and Social Science* 561 (1): 111–26.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://cloud.r-project.org/web/packages/patchwork/index.html.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Readr: Read Rectangular Text Data. 2017.

Ristroph, Alice. 2017. "The Constitution of Police Violence." *UCLA L. Rev.* 64: 1182.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*.