Table of contents

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. This has underscored the urgent need for systematic analysis to inform policies and training within law enforcement agencies. 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. 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. 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 ultimately predicts that specific genders and races have no influence to use violence during law enforcement actions.

```
## we use R [@citeR] for all data wrangling and analysis and R packages tidyverse [@tidy], g
```

data

download data

The source data for this article comes from Open Data Toronto 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.

```
# A tibble: 6 x 6
  `_id` Type_of_Incident           Gender_of_People_Inv~1 Perceived_Race_of_Pe~2
  <int> <chr>                      <chr>                  <chr>
1     1 Reported Use of Force Inc~ Women                  "Black "
2     2 Reported Use of Force Inc~ Women                  "East/Southeast Asian~
3     3 Reported Use of Force Inc~ Women                  "Indigenous "
4     4 Reported Use of Force Inc~ Women                  "Latino "
5     5 Reported Use of Force Inc~ Women                  "Middle Eastern "
6     6 Reported Use of Force Inc~ Women                  "South Asian "
# i abbreviated names: 1: Gender_of_People_Involved,
#   2: Perceived_Race_of_People_Involv
# i 2 more variables: Incident_Count <int>, ObjectId <int>
```

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

```
  Gender_of_People_Involved Perceived_Race_of_People_Involv Incident_Count
1                       Men                           White           1615
2                     Women                          Latino          15486
3         Mixed Gender Group                 Middle Eastern           4874
4         Mixed Gender Group                          White          21320
5                       Men             Multiple race group          20558
6                       Men                 Middle Eastern           7245
```

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.

```
# A tibble: 8 x 2
  Perceived_Race_of_People_Involv Incident_Count
  <fct>                                    <dbl>
1 Black                                    19594
2 East/Southeast Asian                      5848
3 Indigenous                                1995
4 Latino                                    1987
5 Middle Eastern                            4080
6 Multiple race group                      13987
7 South Asian                               4700
8 White                                    35218
```

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.

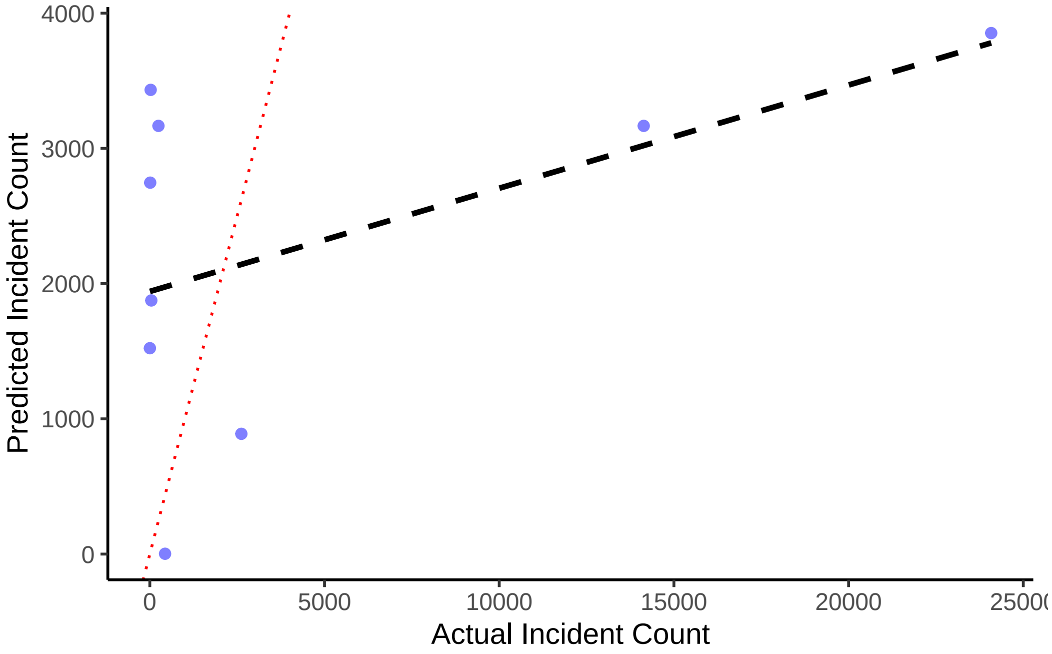
Mean Squared Error	Root Mean Squared Error
62891760	7930.43

analize the model

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.



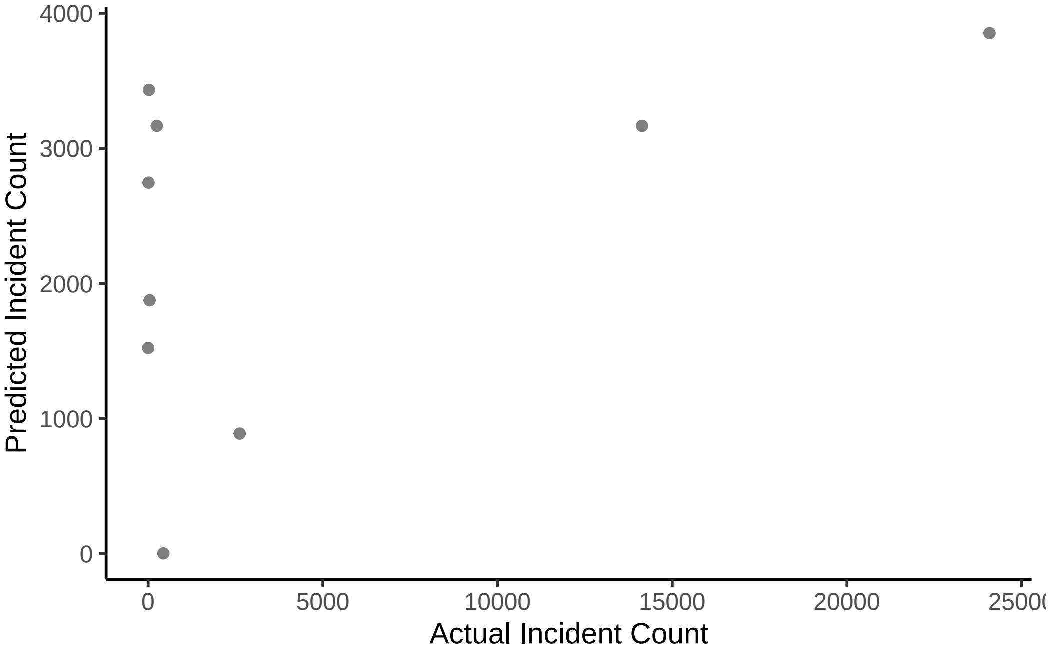
graph analize

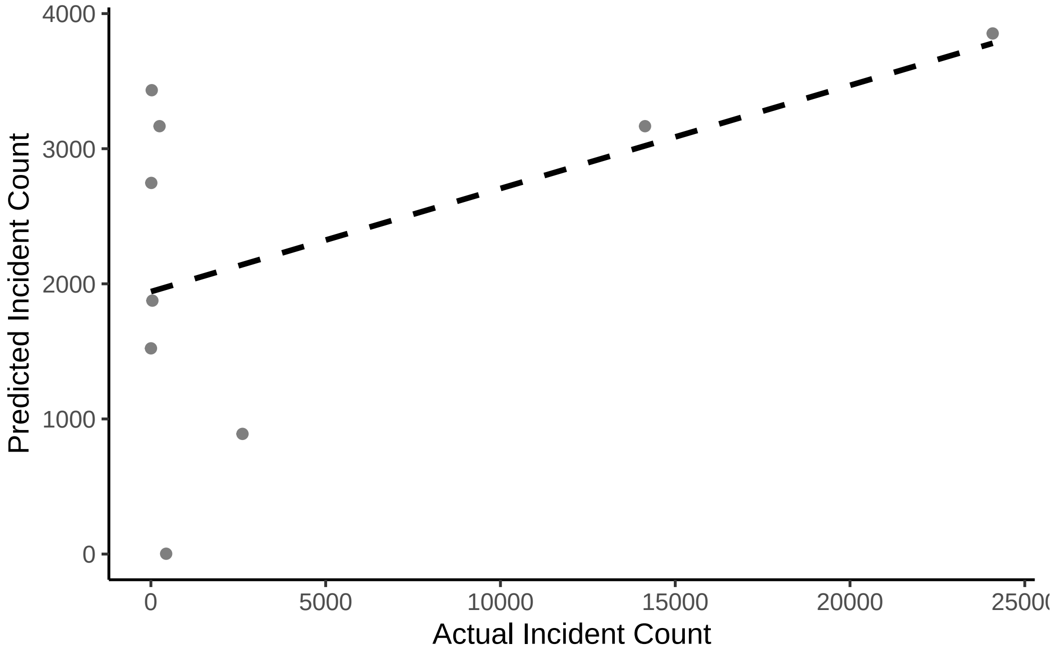
This graph illustrates the relationship between the actual and predicted incident counts as derived from the linear model. The x-axis denotes the actual incident counts, and the y-axis represents the predicted counts based on the model. The dashed line represents the ideal situation where the predictions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

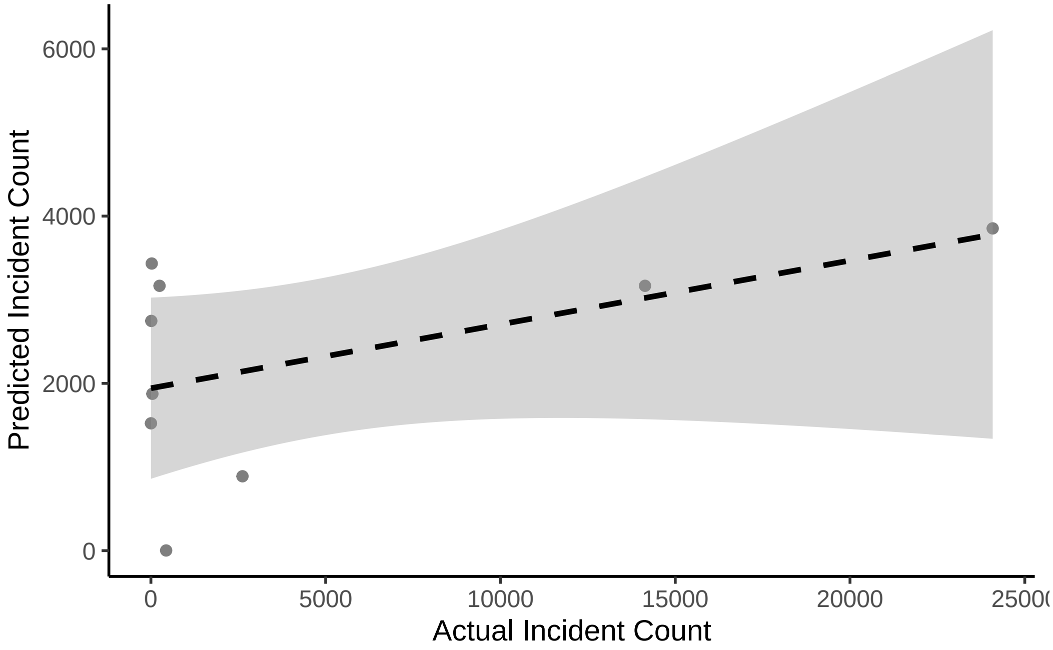
From the graph, we can observe that the points do not align perfectly with the dashed line, indicating some level of prediction error. The shaded area around the dashed line represents the confidence interval, providing a visual representation of the uncertainty in the predictions. As the actual incident 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 underpredict 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.




Predicted Incident Count
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
Actual Incident Count


Predicted Incident Count
6000
4000
2000
0
0
5000
10000
15000
20000
25000
Actual Incident Count

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.

```
#| echo: false
#| warning: false
# If you want to compare multiple models, you can pass them as a list
```

	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

```r
# For example, if you had another model called model2, you could do:
models_list <- list(Model1 = model) #, Model2 = model2)
modelsummary(models_list)
```

discussion