Analysing a Linear Model on the Annual Statistical Report on Homicides in Toronto between 2004 to 2023*

My subtitle if needed

Sarah Lee

March 16, 2024

This paper analysizes the Anuual Statistical Report on Homicides in Toronto between 2004 to 2023 retrieved fromthe OpenDataToronto portal. On the whole, the most noticable factor is shootings being the most common homicide method. On top of that, it is shown that weekends are the most prominant days of when Homicide rates are high. The Poisson and Negative Binomial Models are used in this paper, where the Negative Binomial Model is a more accurate choice. The results of the model shows us that the data given does not have strong predictive power indicated by the R^2 and R^2 Adj. are very low. Hence there would need further interpretation to understand the relationship between the variables

1 Introduction

Homicide rates serve as a crucial indicator of the public's safety as it reveals trends and the distribution of the rates throughout the years. Studying statistics on homicide can motivate the act of public safety and regulations based on the given results and history. This paper explores the distribution of homicide rates between 2004 and 2023 in Toronto along with its predominant methods. The data set on the "Police Annual Statistical Report - Homicides" is retrieved from the Open Data Toronto (Gelfand 2022) portal website given by Toronto Police reports (Services 2024). This resource provides a detailed overviews of the city's homicide counts distributed across various expenditure categories. This paper explores the homicide population of the leading methods and days of homicides in Toronto over the years between 2003 and 2023. By identifying the methods of homicides, governments can prioritize research on these methods to decrease the use of them, by enacting law regulations as such. Additionally,

*Code and data are available at: https://github.com/sarahhhh02/murrumbidgee_paper.git.

police officials by foreseeing the potential days of the week homicides are most likely to occur, they can be extra cautious on those days.

The data is given by a csv. file that includes the categories of location, year, day, month, day of the week, police division, and its unique ID case. The homicide counts are given by the method of homicides, distributed by "shooting", "stabbing", and "other". This paper will mainly focus on the methods of homicides, days of the week, and the years.

With this data set, I plan to use the R programming language (R Core Team 2023) with its relevant tools like Tidyverse (Wickham et al. 2019), janitor (Firke 2023), knitr (Xie 2014), rstanarm (Goodrich et al. 2022) and dplyr (Wickham et al. 2023). With the use of this language, I will build linear models for the variables homicide type and the day of the week and consider their results.

2 Data

The major trends and patterns will be analyzed using the tool from ggplot2 (Wickham 2016) that will graph the needed information. I will also make use of the tool knitr (Xie 2014) to construct tables to give a generalized view on what we are looking for in this paper.The original data set that was retrieved from the Open Data Toronto portal (Gelfand 2022) which includes all cases of homicides from 2004 to 2023. Taking this data set, I cleaned the data into tree separate csv files. The file homicides_clean1 contains the cleaned file of renaming and only taking the columns needed for this paper. Table 1 demonstrates the cleaned data set and shown below Table 1 in Figure 1 is a linear regression visual of the number of homicides per year through 2004 and 2023.

Table 1: Total Homicide Cases per Year

Year	Total Homicides
2004	64
2005	80
2006	70
2007	86
2008	70
2009	62
2010	65
2011	51
2012	57
2013	57
2014	58
2015	59
2016	75

Year	Total Homicides
2017	65
2018	98
2019	79
2020	71
2021	85
2022	71
2023	73

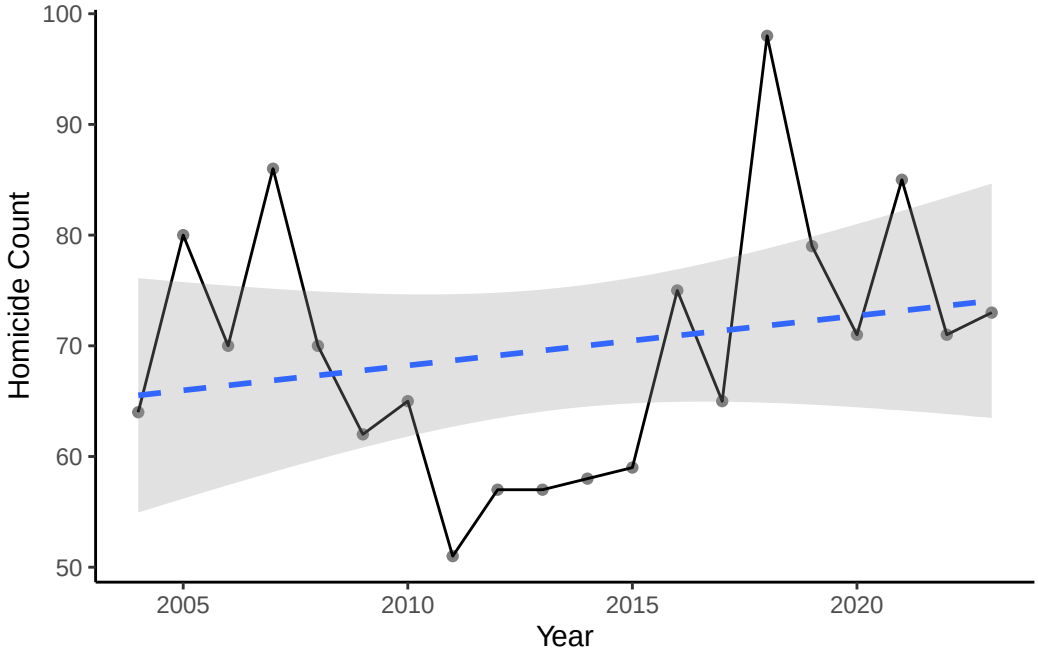


Figure 1: Homicides Counts per Year from 2004 to 2023

With the homicides_clean1 data, I combined the year and homicide types to obtain an integer column consisting of the total homicide counts by each homicide type and year, saving this file into homicides_clean2. This new data contains how many counts of homicide there are for each homicide method along with its year. This will give us an overview on which methods of homicides were more predominant in the years. Shown in Figure 2 we are able to see a side to side comparison of the line graph representation of each method of homicides, "shooting", "stabbing", and "other". Just looking at this visualization shows us how much more shooting is the more predominant method.

Lastly, in another file, homicides_clean3, I combined the year and day of week to again obtain an integer column to see the total homicide counts by year and day of week. Shown below in Figure 3, you can see that the days that are on average predominate in the higher counts of

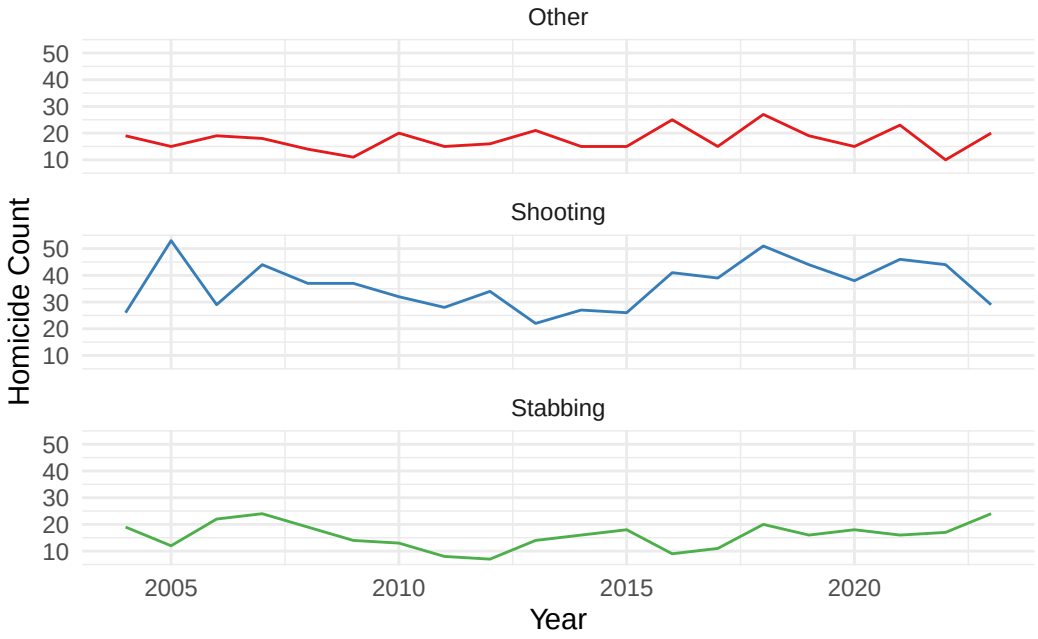


Figure 2: Data Plot of Homicides Counts per Year by Homicide Type

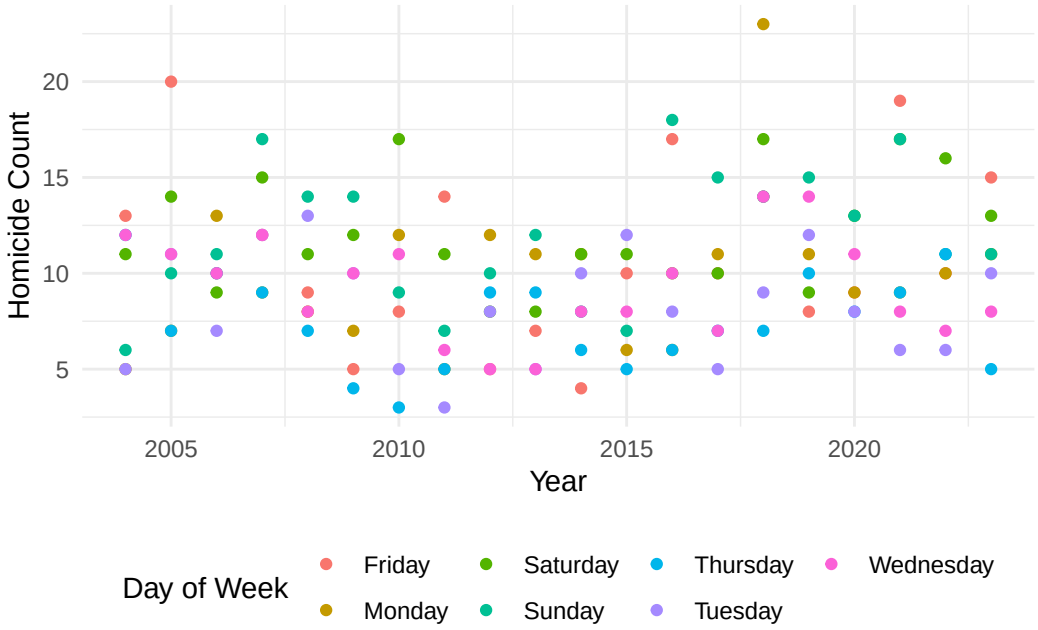


Figure 3: Data Plot of Homicide Counts per Year by Day of Week

homicide is during the weekends, “Friday”, “Saturday”, and “Sunday”. But we can also see that the highest number of counts of homicide is actually on a “Monday”. Hence with this information, we are not able to fully capture the essence of a trend or pattern quite yet.

3 Model

In this data set, the dependent variable is the count of homicides which is a non-negative integer, in this case form the cleaned data sets we have count of homicides by type and day of week. On the other hand, the independent variable is the year which is the linear model that is representing a continuous variable. Based on these two information, the appropriate model for this would be either a Poisson Regression or a Negative Binomial Regression.

Both the Poisson and Negative Binomial Regressions are used when assumption of the constant variance in linear regression is violated. Poisson Regression specifically assumes the equality of the mean and variance of the dependent variable are equal. While the Negative Binomial Regression on the other hand allows the variance to exceed the mean.

3.1 Model set-up

The Poisson model is given by

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
log(\lambda_i) &= \beta_0 + \beta_1 \cdot x_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
& && (5)
\end{aligned}
$$

and the Negative Binomial model is given by

$$
\begin{aligned}
y_i|\lambda_i, \theta &\sim \text{NegativeBinomial}(\mu_i, \theta) && (6)\\
log(\mu_i) &= \beta_0 + \beta_1 \cdot x_i && (7)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (8)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (9)\\
& && (10)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

The Negative Binomial model is a more accurate representation of the data compared to the Poisson model since Poisson Regression is more prone to errors. If we collect data on the mean and variance from the data plots from earlier. It is shown in Table 2, Table 3 and Table 4 that all of their variances exceed their means.

Table 2: Comparison of Mean and Variance of Total Homicides in from 2004 to 2023

Mean	Variance
69.8	136.5895

Table 3: Comparison of Mean and Variance of Total Homicides by Type

Mean	Variance
23.26667	126.8768

Table 4: Comparison of Mean and Variance of Total Homicides by Day of Week

Mean	Variance
9.971429	13.61069

Hence based on these comparisons, it is evident that the model used is correct for this data set.

4 Results

Our results for the data on homicides per year based on the homicide type are summarized in Table 5. The results on homicides per year based on the day of the week are summarized in Table 6.

Below are the Linear Regression models respectively related with Table 5 and Table 6.

Table 5: Explanatory models of Homicides per Year based on Homicide Type

	First model
(Intercept)	24.51
	(35.50)
homicide_typeShooting	−0.12
	(30.90)
homicide_typeStabbing	1.04
	(30.73)
count_type	−0.03
	(1.26)
Num.Obs.	60
R2	0.000
R2 Adj.	−1.000
Log.Lik.	−586.394
ELPD	−587.2
ELPD s.e.	0.1
LOOIC	1174.3
LOOIC s.e.	0.2
WAIC	1174.3
RMSE	1989.38

Table 6: Explanatory models of Homicides per Year based on the Day of Week

	second model
(Intercept)	37.14
	(45.03)
dayMonday	−0.34
	(41.46)
daySaturday	0.32
	(39.78)
daySunday	0.50
	(42.31)
dayThursday	−1.69
	(41.03)
dayTuesday	−0.19
	(39.41)
dayWednesday	0.36
	(41.31)
count_day	0.07
	(3.94)
Num.Obs.	140
R2	0.001
R2 Adj.	−1.000
Log.Lik.	−1302.944
ELPD	−1303.7
ELPD s.e.	0.1
LOOIC	2607.4
LOOIC s.e.	0.2
WAIC	2607.4
RMSE	1975.84

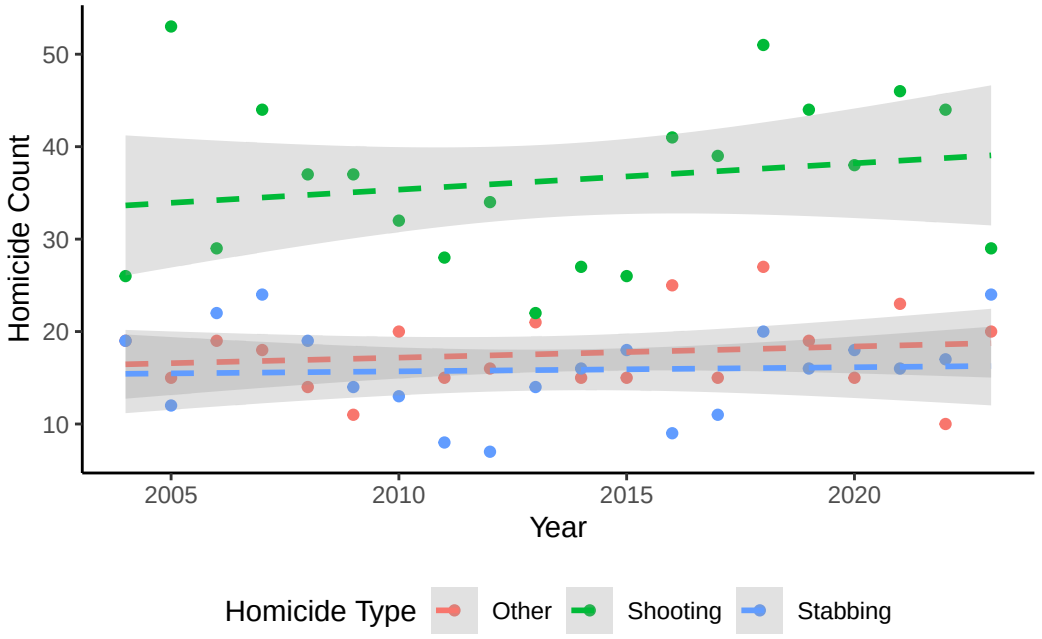


Figure 4: Linear regression with simulated data on the number of homicides per year, depending on the homicide type

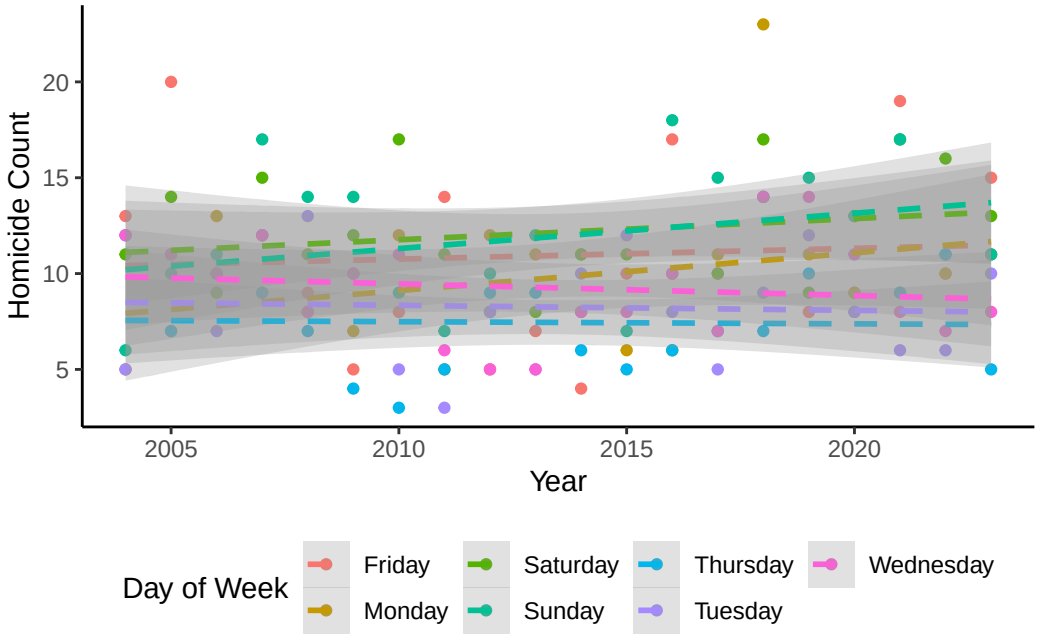


Figure 5: Linear regression with simulated data on the number of homicides per year, depending on the day of week

5 Discussion

5.1 First discussion point

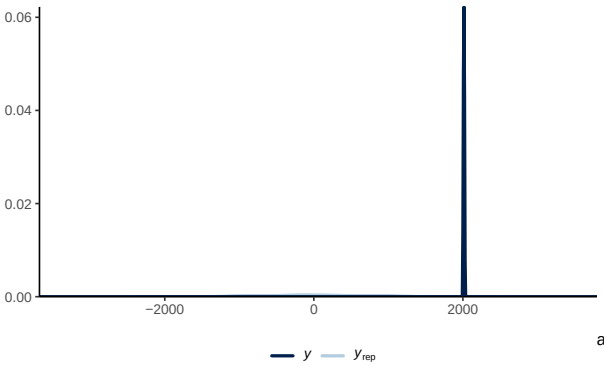
5.2 Second discussion point

5.3 Third discussion point

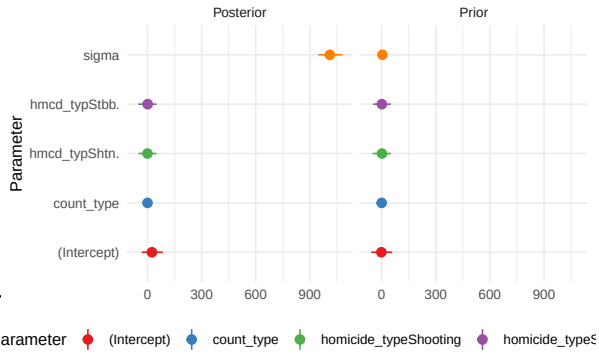
5.4 Weaknesses and next steps

Appendix

.1 Posterior predictive check



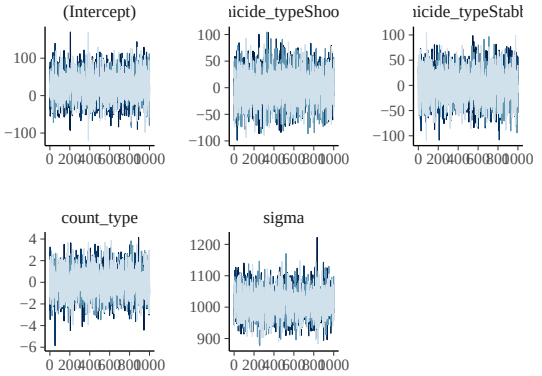
(a) Posterior prediction check



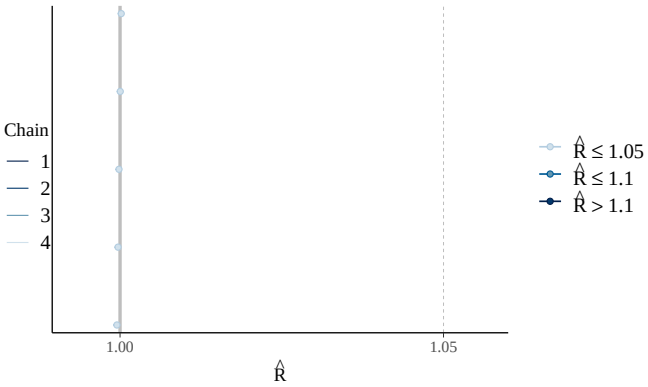
(b) Comparing the posterior with the prior

Figure 6: Examining how the model fits, and is affected by, the data

.2 Diagnostics



(a) Trace plot



(b) Rhat plot

Figure 7: Checking the convergence of the MCMC algorithm

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

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

Services, Toronto Police. 2024. *Police Annual Statistical Report - Homicides*. https://open.toronto.ca/dataset/police-annual-statistical-report-homicide/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.