My title*

My subtitle if needed

Kevin Roe

December 1, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/Kanghyunroe/traffic_collisions/tree/main.

1 Introduction

Automobile fatalities are a leading cause of death worldwide, posing a public health and safety concern for urban cities. For example, in 2024, thirty people have died on Toronto's roadways so far, which is a 20% increase from last year (insert citation, CBC). However, new headlines, such as CBCs, highlight that Motor Vehicle Collisions (MVC) tend to either be generalized as a summary statistic or typically fatal events are over analyzed to the extent that environmental factors surrounding the crash are ignored (CLARIFY). Moreover, general environmental factors such as the time of the crash or if a bicycle was involved in the accident are important to understand what increases the likelihood of a fatality occurring in an MVC. The use of statistical modeling on increasingly available vehicle collision data presents an opportunity to develop a nuanced understanding of what factors increases the likelihood for a fatality to occur in an accident. This paper uses the Toronto Police Service's Annual Statistical Report from Open Data Toronto to analyze what factors are most responsible in predicting if a fatality occurs in a car crash.

The estimand of interest is the log-adjusted probability of a fatality occuring in a car crash. Specifically, we aim to quanitfy how specific environmental factors increase or decrease the likelihood of a fatality. By applying inferential analysis through Bayesian linear models, we assess not only the magnitude of these effects, but their underlying uncertainties (EDIT).

what was found (FINISH PARAGRAPH)

The paper is not only important from a public health perspective, but the paper also has policy development implications. Road safety and reducing fatal MVCs are a critical agenda item of any municipal government. The paper informs what factors increase the likelihood of death in an MVC, which informs policymakers' focus for relevant policy design.

The remainder of this paper is structured as follows: Section 2 describes the dataset and methodology and **?@sec-model** exhibits the use of inferential models. **?@sec-results** presents the results of the analysis, detailing the observed relationships between likelihood of death and various circumstantial factors. **?@sec-discussion** discusses the broader implications and limitations of our findings.@sec-appendix presents a detailed idealized methodology to improve data collection, and additional model summary and diagnostic information.

2 Data

2.1 Overview

This dataset, "Police Annual Statistical Report - Traffic Collisions", was published and refreshed on October 21st, 2024, by the Toronto Police Service [insert citation]. The Toronto Police Service publishes various datasets on public safety and crime to inform the public about safety issues (**annual_statistics_report?**). Data on traffic collisions is included in the Toronto Police Service's Annual Statistical Report, which also covers reported crimes, search of persons, firearms, and the Police Service's budget (**annual_statistics_report?**). The data is collected using historical Motor Vehicle Collisions and classifies them into the following categories: * Property Damage (PD) Collisions * Fail to Remain (FTR) Collisions, or commonly known as hit-and-run accidents * Injury Collisions * Fatalities

Following the Municipal Freedom of Information and Protection of Privacy Act, the Toronto Police Service ensures to protect the privacy of individuals involved in the reported crimes when publishing the data. The dataset is updated annually, is open data, and can be used if an attribution statement **?@sec-appendix-attribution** and is properly cited (**tphlicense?**).

There is an alternative dataset from the Toronto Police Service called "Motor Vehicle Collisions involving Killed or Seriously Injured Persons" (CITE). Unlike the alternative dataset, this paper's dataset focuses on all collisions, instead of only focusing on ones where someone was either killed or seriously injured. While the alternative dataset has more explanatory variables simply because more data is collected when someone dies or is seriously injured, this paper's aims to generalize if a fatality is more likely to occur based on the general circumstances surrounding a crash, such as the time of day or if property damage occurred. Thus, we ended up not going with the alternative dataset for this paper, but there are variables in the alternative dataset that may motivate future research on this subject. (EDIT TO MAKE MORE CLEAR)

The paper uses the R programming language (R Core Team 2023) to analyze the dataset. The tidyverse package was used to simulate the dataset. Also, the tidyverse (**citetidyverse?**), arrow [CITE] and opendatatoronto (**citeopendatatoronto?**) packages were used to download the Victims of Crime dataset. Then, the tidyverse (**citetidyverse?**) package was used to clean the raw dataset and generate tests. The testthat package [CITE THIS] was used to create tests for our cleaned dataset. Rstanarm [CITE], Arrow [CITE], and bayestestR [CITE] were used to

create and test the model. Finally, ggplot2 (**citeggplot2?**), tidyverse (**citetidyverse?**), knitr (**citeknitr?**) and scales (**citescales?**) packages were used to create the tables and graphs to display the results. [edit this paragraph]

2.2 Measurement

Transforming a real-life Motor Vehicle Collision to an entry in the dataset is a well-documented process by the Toronto Police Service. For insurance purposes, the Toronto Police Service requires drivers to fill out the Motor Vehicle Collision Report for any collisions that occur in Toronto (CITE). Drivers required to fill out a motor vehicle collision report if the combined damage is more than $2000, if someone is injured, if a criminal act such as a DUI occurs, or if a pedestrian is involved in the accident (CITE). These reports are retained for six years by the Toronto Police Service, with the exception of collisions resulting in a fatality, which are retained indefinitely. The form ensures documentation of collision characteristics, location, road condition, and the extent of damages, systematically recording the characteristics of each crash for further criminal investigation and data analysis.

For every collision, basic facts such as the location, time, and date of the collision is recorded through the Motor Vehicle Collision Report. Majority if not all the factors recorded in the dataset are all objective measurements regarding the specific details such as if a motorcycle was involved in the collision or if the collision resulted in property damage. All of these details are recorded in the Motor Vehicle Collision Report for all vehicular collisions and are entered into the dataset. However, while the Motor Vehicle Collision Report logs characteristics such as environmental conditions, alcohol involvement, or fatigue, the data set does not include them due to inconsistent data measurement techniques. Moreover, personal details such as the driver's age are not included to protect the driver's privacy.

2.3 Outcome variables

The main outcome variable records the number of fatalities for each car crash. However, because we are more interested in predicting the probability that a fatality occurs than the number of fatalities, we transformed the variable that distinguishes collisions between if the collision resulted in any fatalities and those without fatalities. In the raw dataset, if there were no fatalities, the entry was recorded as NA, but if there were fatalities, then the number of fatalities were recorded. However, we transformed the dataset that all if a fatality occurred then fatalities indicates `1` and if there were no fatalities then the fatalities column records a '0'

However, after the transformation, the outcome variable now takes on binary values and the distribution and summary statistics are shown below:

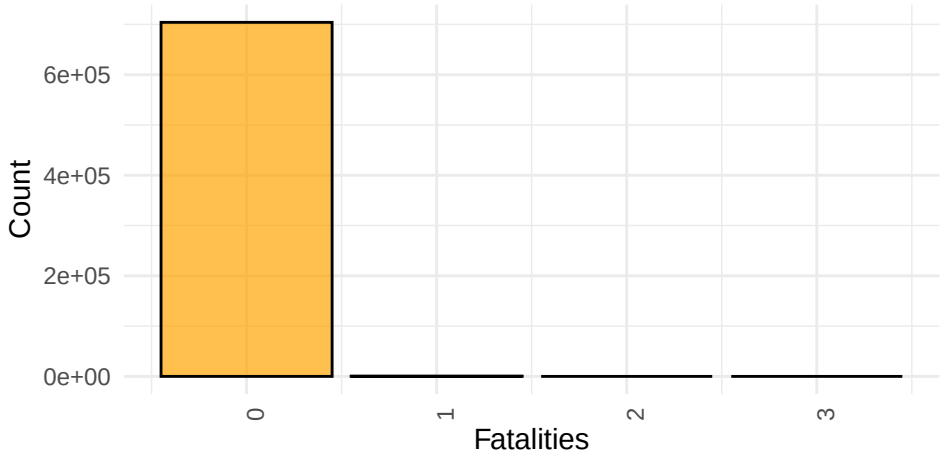

Figure 1: Distribution of Fatalities in the Raw Dataset

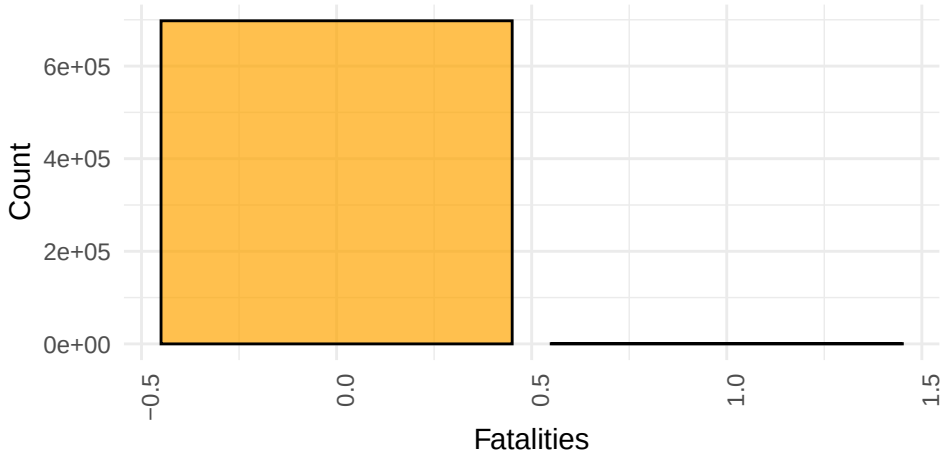

Figure 2: Distribution of Fatalities in the Cleaned Dataset

Table 1: Number of unique high-quality polling organizations

fatalities	Fatalities
0	697852
1	606

2.4 Predictor variables

The variables of interest in the paper are CrimeType (named SUBTYPE in the original dataset), which categories the type of crime into the four areas of Assault, Sexual Violation, Robbery and Other; AssaultType (named ASSAULT_SUBTYPE in the original dataset), which specifies assault on peace officers into the subtypes noted in **?@sec-introduction**; Sex (named SEX in the original dataset) - which is broken down into Male, Female and Unknown, where Unknown means the victim's sex is not known by the Toronto Police Service; and Count (named COUNT in the original dataset), which counts the number of identified victims who share the same demographic characteristics previously. Each entry in the dataset does not represent a unique person but notes the number of people who share the same characteristics, such as Sex and CrimeType. Other characteristics such as age group, reported year and age cohort were not variables of interest for the study and were removed from the cleaned dataset. Using knitr (**citeknitr?**), the first 10 lines of the cleaned dataset is shown in **?@sec-appendix-sample** under **?@tbl-head**. In addition, summaries of the Sex and Year variables are displayed in **?@sec-appendix-summary_statistics**.

2.4.1 Hour

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

2.4.2 Type of Crash

2.4.3 Automobile

2.4.4 Motorcycle

2.4.5 Passenger

2.4.6 Pedestrian

2.4.7 Police Division

Fixed effects

2.4.8 Year

Fixed effects

2.4.9 Interaction Terms

Why you have interaction between hour and all these things

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

Table 2: Explanatory models of flight time based on wing width and wing length

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

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

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 2.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In we implement a posterior predictive check. This shows...

In we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

is a trace plot. It shows... This suggests...

is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algo-
rithm

References

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/.