When Risk Hits the Road: Does Danger Drive a Return?*

Motor Theft and Collisions in Toronto Are Driven by [Neighborhood Clustering] and [Environmental Conditions]

Yingke He

November 30, 2024

This study examines motor vehicle theft and traffic collisions in Toronto, focusing on spatial and temporal patterns, recovery outcomes, and contributing factors. [The analysis reveals that thefts are geographically clustered, with recovery rates differing by location type, while collisions are more frequent under poor visibility and adverse road conditions.] These findings underscore the need for targeted interventions, including enhanced surveillance in high-theft areas, infrastructure improvements for road safety, and optimized strategies for vehicle recovery. Such measures are critical for fostering safer and more secure urban mobility in Toronto.

Table of contents

*Code and data are available at: https://github.com/ohyykk/Toronto_Motor_Viehicle/tree/main.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (**shelter?**)…. Following Alexander (2023), we consider…

Overview text

2.2 Source

2.3 Data Measurement and Limitations

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.4 Outcome Variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from (**palmerpenguins?**).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.5 Predictor Variables

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.

3 Model

The main idea of this composite risk score model is to calculate an index of risk for owning and riding a motorbike. The modeling strategy has two primary objectives. The first objective is to estimate the likelihood of theft (P(Theft)) and collision (P(Collision)) under diverse environmental and situational conditions using logistic regression models. The second objective is to combine these probabilities into a unified Risk Index, providing actionable insights into the risk for owning and riding a motorbike

The models were run in R (R Core Team 2023) using the `stats` package to implement logistic regression models predicting theft and collision probabilities. The theft model incorporates predictor variables such as neighborhood ID, premises type, report hour, location type, and report day, while the collision model includes neighborhood ID, road surface condition, lighting condition, and traffic control.

Both models estimate the log-odds of their respective probabilities using the predictors, enabling reliable predictions under varying environmental and situational conditions. Model diagnostics and background details, including variable descriptions and performance metrics, are included in Appendix B.

3.1 Model Set-up

The logistic regression model for theft probability takes the form of the following equation:

$$\log\left(\frac{P(\text{Theft})}{1-P(\text{Theft})}\right) = \beta_0 + \beta_1 \cdot \text{HOOD}_{158} + \beta_2 \cdot \text{Premises Type} + \beta_3 \cdot \text{Report Hour} \\ + \beta_4 \cdot \text{Location Type} + \beta_5 \cdot \text{Report Day} + \epsilon$$

This model utilizes the following predictor variables:

- **Neighborhood ID** (`hood_158`): Unique identifier for the neighborhood.
- **Premises Type** (`premises_type`): Type of premises where the theft occurred like parking lot, garage
- **Report Hour** (`REPORT_HOUR`): Hour of the day the theft was reported.
- **Location Type** (`location`):Location of the theft like indoors and outdoors.
- **Report Day** (`REPORT_DOW`): Day of the Week Offence was Reported.

The logistic regression model for collision probability takes the form of the following equation:

$$\log\left(\frac{P(\text{Collision})}{1-P(\text{Collision})}\right) = \beta_0 + \beta_1 \cdot \text{HOOD}_{158} + \beta_2 \cdot \text{Road Surface Condition} + \beta_3 \cdot \text{Light} \\ + \beta_4 \cdot \text{Traffic Control} + \epsilon$$

This model utilizes the following predictor variables:

- **Neighborhood ID** (`hood_158`): Unique identifier for the neighborhood.
- **Road Surface Condition** (`road_conditions`): Road surface condition like wet and dry.
- **Lighting Condition** (`lighting_conditions`): Lighting conditions at the time of the collision like dark or clear.
- **Traffic Control** (`traffic_control`):Type of traffic control present like stop signs or traffic signals After fitting both models, combine the probabilities into a weighted risk index:

$$\text{Risk Index} = w_1 \cdot P(\text{Theft}) + w_2 \cdot P(\text{Collision})$$

where weights $w_1$and w_2 are adjust based on the relative importance of theft and collision risks.

3.2 Model justification

Logistic regression was selected for this analysis due to its simplicity and effectiveness in estimating probabilities for binary outcomes, such as theft and collision. This statistical method is well-suited for modeling the log-odds of an event's occurrence as a linear combination of predictor variables, providing a clear framework for understanding the relationships between independent variables and the dependent outcome. Its ability to estimate odds ratios allows for an intuitive interpretation of the impact of predictor variables, making it particularly effective for identifying and quantifying risks in diverse scenarios.

This model's strengths include its flexibility in handling both categorical and continuous predictors, enabling the integration of a wide range of variables such as neighborhood characteristics, premises type, lighting conditions, road surface conditions, time of day, and environmental conditions. Additionally, logistic regression's ability to include interaction terms and manage potential multicollinearity enhances its effectiveness in capturing complex relationships between predictors and the likelihood of thefts and collisions. Utilizing these features enables the model to deliver a thorough evaluation of the factors affecting theft and collision probabilities, addressing the first objectives of the modelling strategy.

After estimating the probabilities of theft and collision, these probabilities are combined into a composite Risk Index using a weighted formula. This Risk Index offers a unified measure of risk, allowing for comparative evaluations of the relative safety of motorbike ownership and usage across varying contexts.

The modeling approach is justified by logistic regression's ability to deliver clear, interpretable, and statistically reliable risk estimates. Its statistical efficiency, with its capacity to manage complex, large datasets, makes it an ideal choice for accurately predicting theft and collision risk, while the composite Risk Index facilitates a clearer understanding of the risks associated with motorbike ownership and usage.

3.3 Model Assumtions and Validations

To validate the use of logistic regression models, the four main assumptions: binary nature of the outcome, linearity of the logit, indipendence of observations, and lack of multicollinearity will be assessed (Kononen, Flannagan, and Wang 2011). To validate the use of a composite Risk Index, the correlation of the index with observed outcomes will be assess using a Pearson correlation test.

3.3.1 Logistic Regression Models

3.3.1.1 Binary Nature of the Outcome

A fundamental assumption of logistic regression is that the response variable is binary or dichotomous, meaning it can take on only two possible outcomes (Nick and Campbell 2007). This assumption is satisfied in both models. In the theft model, the response variable indicates whether a theft occurred, with a value of 1 representing the occurrence of theft and 0 representing its absence. Similarly, in the collision model, the response variable denotes whether a collision occurred, where 1 represents the occurrence of a collision and 0 indicates no collision.

The logistic regression model does not directly predict 1 or 0. Instead, it estimates the probability that the response variable equals 1, given the predictor variables. For instance, the theft model outputs the probability of theft occurring (P(Theft)), and the collision model outputs the probability of a collision occurring (P(Collision)). These probabilities range between 0 and 1, allowing for detailed predictions of the likelihood of the respective events.

3.3.1.2 Independence of Observations

A second assumption of logistic regression is the independence of observations, meaning that all outcomes within the sample must be separate from one another (Stoltzfus 2011). This implies that the data should not contain repeated measurements or outcomes that are inherently correlated.

For this analysis, the data has been checked to ensure that it contains no duplicate measures, with each observation corresponding to a distinct event, such as a theft or collision.

However, it is acknowledged that the data includes groupings by neighborhood, as indicated by the variable Neighborhood ID (`hood_158`). Neighborhood groupings may introduce some correlation due to shared environmental or socio-demographic factors. Despite this, the variable Neighborhood ID is included in the models as a predictor to account for such shared characteristics explicitly. Additionally, the other predictors, such as Road Surface Condition, Lighting Condition, and Traffic Control for the collision model, and Premises Type, Report Hour, Location Type, and Report Day for the theft model, represent independent situational or environmental variables tied to specific events.

Given the heterogeneity across neighborhoods and the distinct nature of the predictors, any residual correlation within groupings is expected to be minimal and not comparable to the stronger dependency seen in matched data, such as multiple observations from the same household or individual. Accordingly, it is believed that the independence assumption of logistic regression is sufficiently satisfied for this analysis.

3.3.1.3 Linear Relationship in the Log-Odds

3.4 “‘{r}

4 Load necessary libraries

library(dplyr) library(here)

dataset <- read_csv(here(“data”, “02-analysis_data”, “cleaned_traffic_data.csv”))

5 Fit the logistic regression model for theft

model <- glm(outcome ~ REPORT_HOUR + premises_type + location + REPORT_DOW, family = binomial, data = dataset)

6 Create bins for REPORT_HOUR

dataset$hour_bin < -cut(dataset$REPORT_HOUR, breaks = 10) # Divide into 10 bins

7 Calculate mean REPORT_HOUR and predicted probabilities for each bin

bin_summary <- dataset %>% group_by(hour_bin) %>% summarise(mean_hour = mean(REPORT_HOUR, na.rm = TRUE), mean_prob = mean(predict(model, type = “response”), na.rm = TRUE))

8 Plot the relationship between mean REPORT_HOUR and predicted probabilities

plot(bin_summary$mean_hour, bin_summary$mean_prob, xlab = “Mean Report Hour (Binned)”, ylab = “Mean Predicted Probability”, main = “Linearity Check for REPORT_HOUR”, pch = 20, col = “blue”) lines(lowess(bin_summary$mean_hour, bin_summary$mean_prob), col = “red”, lwd = 2)

“‘

8.0.0.1 Absence of Multicollinearity

8.0.1 Composite Risk Index

8.0.1.1 Pearson Correlation Test

9 Results

Our results are summarized in **?@tbl-modelresults**.

9.1 First Result Point

9.2 Second Result Point

9.3 Third Result Point

10 Discussion

10.1 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.

10.2 Second discussion point

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

10.3 Third discussion point

10.4 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 **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

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

Kononen, Douglas W., Carol AC Flannagan, and Stewart C. Wang. 2011. “Identification and Validation of a Logistic Regression Model for Predicting Serious Injuries Associated with Motor Vehicle Crashes.” *Accident Analysis & Prevention* 43 (1): 112–22.

Nick, Todd G., and Kathleen M. Campbell. 2007. “Logistic Regression.” In *Topics in Biostatistics*, 273–301. Springer.

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

Stoltzfus, Jill C. 2011. “Logistic Regression: A Brief Primer.” *Academic Emergency Medicine* 18 (10): 1099–1104.