My title*

My subtitle if needed

First author Another author

November 13, 2024

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

1 Introduction

1.1 Overview

This paper investigates the factors influencing the severity of motor vehicle collisions, with a particular focus on the conditions at the time of each incident. Motor vehicle collisions remain a leading cause of death and severe injury globally, and understanding the circumstances under which these incidents result in severe outcomes is critical for improving road safety. By examining data from the city of Toronto, we aim to identify how driver condition, road surface condition, and lighting condition contribute to the likelihood of a collision resulting in a severe injury or fatality. The findings of this study can inform targeted road safety policies and interventions to mitigate the risk of severe traffic accidents.

1.2 Estimand

The estimand in this paper is the probability of a severe injury (fatal outcome) in motor vehicle collisions, given specific conditions at the time of the incident, including driver condition, road surface, and lighting conditions. Measuring this probability for all collisions is challenging due to limitations in data completeness and accuracy, as not all collisions are reported or consistently documented. Additionally, the dataset contains broad categories such as "Other" for road and driver conditions, which may obscure specific contextual details. To estimate this probability, we employ a logistic regression model on a sample dataset of collisions from Toronto's open data portal, allowing us to approximate the likelihood of severe injuries under varying conditions.

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

1.3 Results Summary

The results of our logistic regression model indicate significant associations between injury severity and certain predictor variables. Notably, impaired driver conditions and adverse lighting environments are associated with an increased likelihood of severe injuries. However, hazardous road surfaces—counterintuitively—appear to be linked with a lower probability of severe injury, which could reflect behavioral adjustments made by drivers under difficult road conditions. This analysis highlights the complex interplay between environmental and human factors in determining collision outcomes, underscoring the need for multifaceted safety measures.

1.4 Why This Study Matters

This study is important for several reasons. First, it provides actionable insights into the specific conditions under which motor vehicle collisions are most likely to result in severe injury or death. This information can guide policymakers and road safety advocates in designing effective interventions. For example, enhanced education campaigns targeting impaired driving, improvements in road lighting infrastructure, and better management of road surfaces in hazardous conditions could collectively contribute to reducing collision severity. Additionally, this study uses a Bayesian logistic regression model, which allows for the incorporation of prior knowledge and provides a robust framework for understanding the factors influencing injury severity. In a broader sense, understanding the determinants of collision severity can support efforts to create safer urban environments, thereby protecting the well-being of residents and reducing the societal costs associated with severe traffic incidents.

1.5 Paper Structure

The remainder of this paper is structured as follows. In Section 2, we present an overview of the dataset used in this study, along with a detailed description of the variables and the data cleaning process. Section 3 outlines the Bayesian logistic regression model applied in our analysis, including the model setup, assumptions, and justification. In **?@sec-result**, we summarize the results of our analysis and interpret the findings within the context of existing literature on road safety. Finally, Section 5 discusses the implications of these findings, acknowledges the limitations of the study, and suggests directions for future research.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

This study utilizes data from the Motor Vehicle Collisions Involving Killed or Seriously Injured Persons dataset, made publicly available by Open Data Toronto. The dataset contains information on traffic incidents in Toronto where individuals were killed or seriously injured, providing key insights into the factors associated with severe collisions. The data, which includes variables related to the time, location, conditions, and parties involved in each collision, serves as an essential resource for understanding patterns in road safety and identifying high-risk factors.

2.2 Measurement

The Motor Vehicle Collisions Involving Killed or Seriously Injured Persons dataset from Open Data Toronto provides critical information on traffic incidents involving fatalities or serious injuries in Toronto since 2006. Compiled and maintained by the Toronto Police Service, the dataset captures a variety of incident-specific metrics, including severity, driver condition, road surface condition, and lighting at the time of each collision. Incident severity is categorized as either fatal or serious injury, while environmental conditions, such as road surface (e.g., dry, wet, icy) and lighting (e.g., artificial or natural light), are recorded qualitatively. Each record also includes a timestamp and a location, though the precise geographic coordinates are adjusted to the nearest road intersection to protect the privacy of those involved.

Despite its value, this dataset has limitations that impact the accuracy and applicability of analyses. Most notably, the location of each incident is deliberately offset, meaning that any geographic analysis may not reflect the exact sites of collisions, particularly at the neighborhood or divisional level. Additionally, the dataset is limited to incidents involving significant injuries or fatalities, meaning it does not account for minor or unreported collisions. This focus on severe incidents may lead to an overrepresentation of high-severity events, which could influence interpretations of broader traffic patterns in Toronto. Furthermore, while the dataset aims to provide timely and complete information, the Toronto Police Service does not guarantee its accuracy, completeness, or timeliness, cautioning against direct comparisons with other sources of traffic or crime data. This dataset predominantly consists of categorical variables describing conditions and severity, with minimal quantitative data aside from incident timestamps, making it a qualitative yet powerful resource for studying serious traffic events in Toronto.x

2.3 Data Cleaning

The raw motor vehicle collision data underwent a comprehensive cleaning process to prepare it for analysis, ensuring the dataset was both relevant and consistent for the study's objectives. Initially, only key columns were selected from the raw dataset, including `ACCNUM` (Accident Number), `DATE`, `TIME`, `ROAD_CLASS` (Road Class), `DISTRICT`, `TRAFFCTL` (Traffic Control Type), `VISIBILITY`, `LIGHT` (Light Condition), `RDSFCOND` (Road Surface Condition), `ACCLOC` (Accident Location), `ACCLASS` (Accident Class), `IMPACTYPE` (Impact Type), `INVTYPE` (Involvement Type), `INVAGE` (Involved Person Age), `INJURY`, and `DRIVCOND` (Driver Condition). This selection ensured that only the most pertinent information was included for examining factors associated with injury severity.

Rows with missing or irrelevant data in critical columns, such as `INJURY`, `DRIVCOND`, `LIGHT`, and `RDSFCOND`, were filtered out to maintain data integrity and focus on valid observations. Injury severity was then simplified by converting it into a binary variable (`INJURY_SEVERE`), where incidents classified as "Fatal" were assigned a value of 1, and all other cases were coded as 0, distinguishing between fatal and non-fatal outcomes.

Categories within certain variables were combined to facilitate a more cohesive analysis. For `DRIVCOND` (Driver Condition), cases indicating impairment due to alcohol or drugs were grouped under "Impaired," while other conditions such as inattentiveness, fatigue, or medical issues were combined under "Distracted/Impaired." The "Normal" condition was retained as is, while all other conditions were grouped under "Other."

Similarly, `RDSFCOND` (Road Surface Condition) categories were simplified. "Good Condition" was designated for cases with no adverse road conditions, while surfaces like ice, snow, or gravel were grouped under "Hazardous Surface." "Wet" was kept as a separate category, and any other conditions were labeled as "Other." For `LIGHT` (Light Condition), natural lighting conditions such as daylight, dawn, and dusk were grouped as "Natural Light," while artificial lighting conditions were grouped as "Artificial Light." All other conditions were categorized as "Other."

After grouping these variables, the categorical groupings were converted to factors to enable effective modeling. The cleaned dataset was then saved as a parquet file, ready for analysis in the next stages of the study.

2.4 Outcome variables

The primary outcome variable in this study is Injury Severity, which categorizes each collision based on the severity of injuries sustained by the individuals involved. Injury severity is coded as a binary variable, where "0" represents non-fatal or minor injuries, and "1" denotes fatalities or serious injuries. This classification allows for an assessment of factors contributing to the likelihood of severe outcomes in motor vehicle collisions.

Table 2: Statistics summary of the cleaned Motor Vehicle Collisions dataset

```
 Injury Severity            Driver Condition       Road Surface Condition
 0:17981         Normal              : 6158   Good Condition   :   29
 1:  976         Distracted/Impaired: 1842   Hazardous Surface:  409
                 Impaired            :  318   Wet              : 3140
                 Other               :10639   Other            :15379
         Light Condition
 Artificial Light: 4057
 Natural Light   :11144
 Other           : 3756
```

By focusing on collisions resulting in significant harm, this outcome variable highlights critical cases of interest for road safety analysis, as severe injuries or fatalities often have lasting impacts on communities and may indicate areas where safety improvements are urgently needed. This binary categorization also simplifies the statistical analysis, facilitating the application of logistic regression models to estimate the probability of severe injuries under varying conditions.

Analyzing injury severity in conjunction with predictor variables such as driver condition, road surface condition, and light condition helps identify the circumstances under which the risk of severe outcomes is heightened. This outcome measure serves as a crucial indicator for assessing and improving traffic safety policies, as it reflects the most serious consequences of road incidents and directs attention to factors that could reduce such outcomes.

Table 1: Preview of the cleaned Motor Vehicle Collisions dataset

Injury Severity	Driver Condition	Road Surface Condition	Light Condition
0	Other	Wet	Other
0	Other	Wet	Other
0	Normal	Wet	Other
0	Other	Wet	Other
0	Other	Wet	Other

2.5 Predictor variables

Figure 1 shows the distribution of injury severity based on driver condition at the time of the collision. The majority of collisions involve drivers categorized as “Other” or “Normal,” with very few incidents involving drivers who were impaired or distracted/impaired. Despite

the low frequency, collisions involving impaired or distracted drivers show a relatively higher proportion of serious injuries (indicated by "1") compared to the "Normal" category. This highlights that impaired or distracted driving is associated with an increased likelihood of severe outcomes, although it is less common overall.

As shown in Figure 2, the road surface condition at the time of the collision also influences injury severity. The majority of incidents occurred under "Other" or "Wet" conditions, with very few taking place on hazardous or good road surfaces. Wet and hazardous surfaces are associated with a slightly higher proportion of severe injuries, emphasizing that adverse road conditions can elevate the risk of serious outcomes in collisions.

Figure 3 illustrates the distribution of injury severity by light condition, indicating that most incidents occur under "Natural Light" conditions, followed by "Artificial Light." Although the majority of collisions occur during natural light conditions, the proportion of severe injuries is slightly higher in cases with "Other" or "Artificial Light." This may suggest that low visibility or artificial lighting conditions marginally contribute to the severity of collisions.

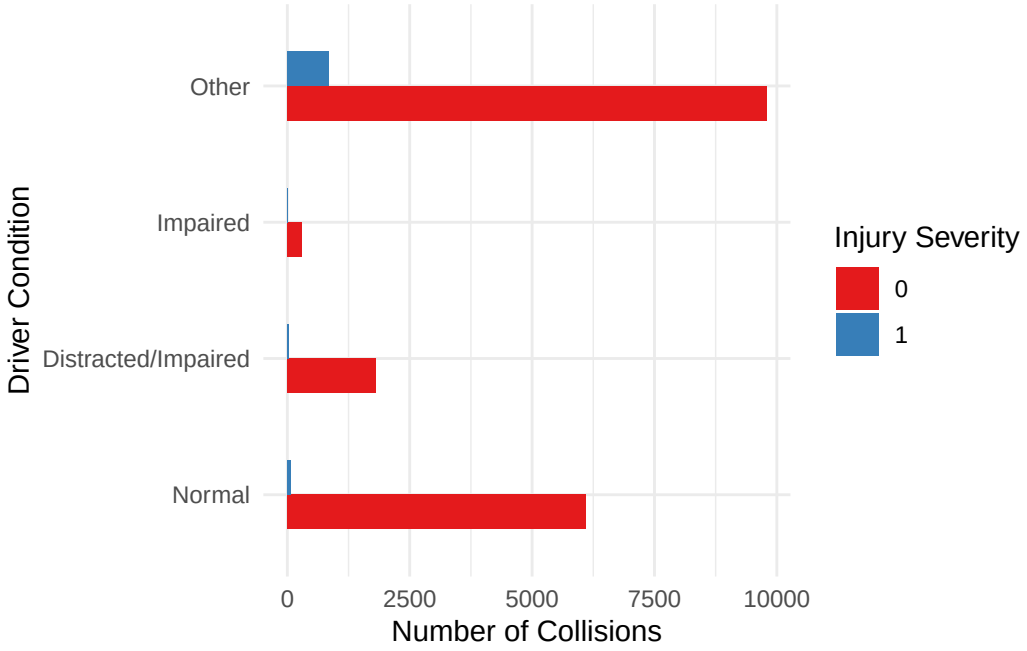


Figure 1: The distribution of injury severity by driver condition

3 Model

In our analysis, we utilized a Bayesian logistic regression model to examine the relationship between injury severity in motor vehicle collisions and three key factors: driver condition, road

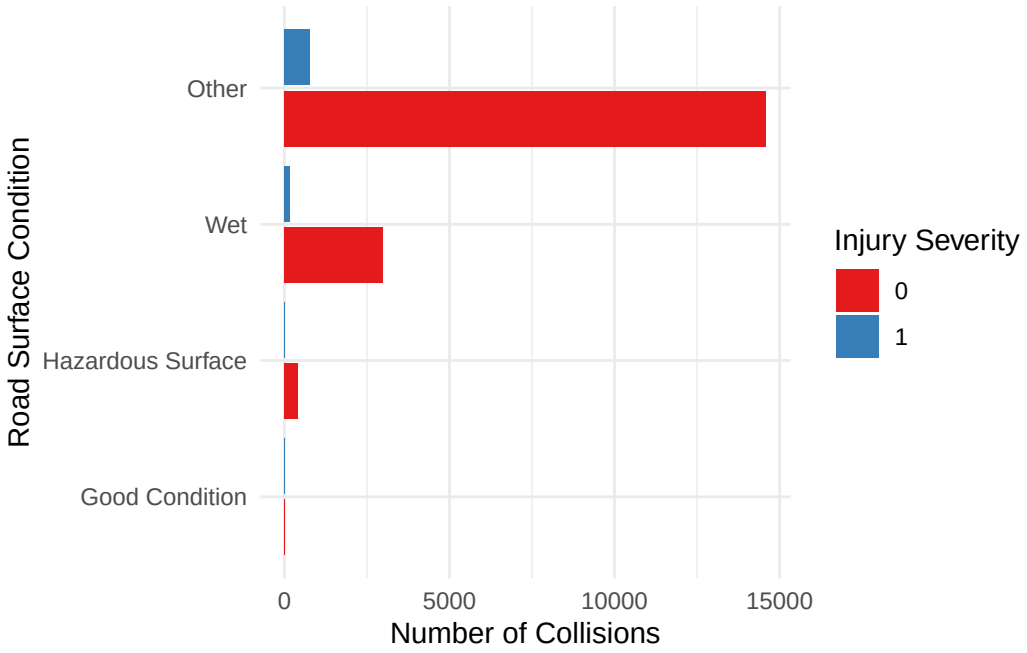


Figure 2: The distribution of injury severity by road surface condition

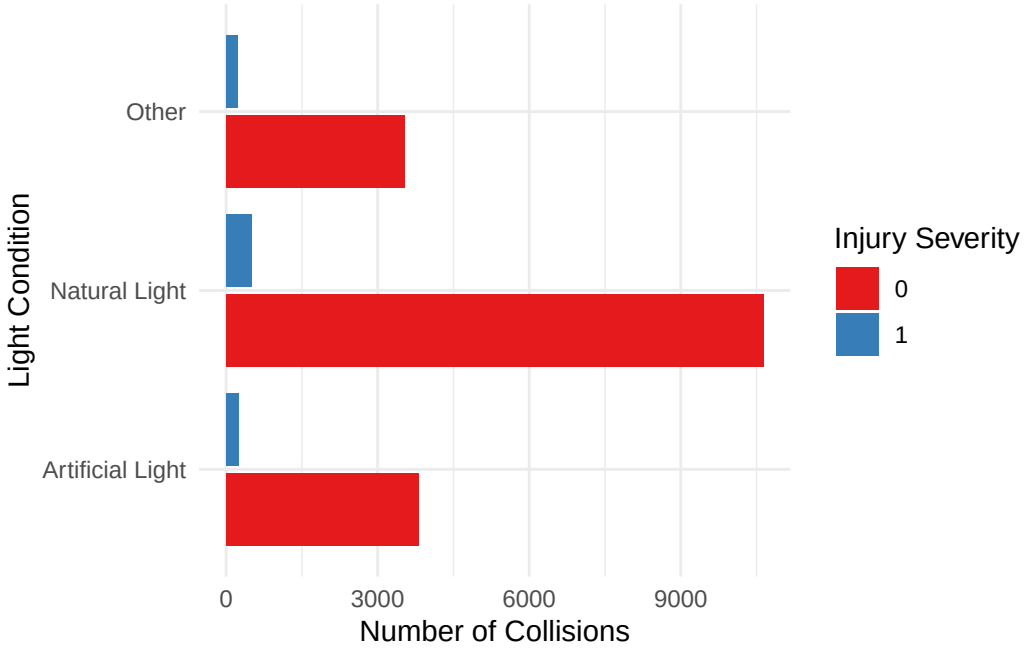


Figure 3: The distribution of injury severity by light condition

surface condition, and lighting condition at the time of the incident. Background details and diagnostics are included in Appendix B.

3.1 Model Set-up

The model is formulated as follows:

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$

$$\text{logit}(\pi_i) = \alpha + \beta_1 \times \text{DRIVCOND\_GROUP}_i + \beta_2 \times \text{RDSFCOND\_GROUP}_i + \beta_3 \times \text{LIGHT\_GROUP}_i \tag{2}$$

$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

In this model, y_i represents the binary outcome variable indicating whether a collision resulted in a severe injury (1 for severe/fatal injuries, 0 for non-fatal injuries). The probability of a severe injury (π_i) is modeled using a logistic link function, which expresses the log-odds of a severe injury as a linear combination of the intercept (α) and the coefficients (β_1, β_2, β_3) associated with the predictor variables: driver condition (`DRIVCOND_GROUP`), road surface condition (`RDSFCOND_GROUP`), and light condition (`LIGHT_GROUP`), respectively.

The intercept (α) and each coefficient (β_1, β_2, β_3) are assigned normal prior distributions with a mean of 0 and a standard deviation of 2.5. These priors serve to regularize the model, ensuring that estimates remain reasonable given the data.

We chose this modeling approach for several reasons. Logistic regression is appropriate for binary outcome variables, making it ideal for analyzing injury severity. Bayesian methods allow us to incorporate prior knowledge and account for uncertainty, leading to more robust parameter estimates. While alternative modeling approaches, such as linear regression, were considered, Bayesian logistic regression was selected due to the binary nature of the outcome variable.

The model was implemented using the `rstanarm` package (**citeRstanarm?**) in R (R Core Team 2023). The posterior distributions of the parameters were estimated using Markov Chain Monte Carlo (MCMC) sampling. To optimize runtime, we randomly sampled 1000 observations from the dataset with a random seed of 215. Diagnostics, including convergence checks and posterior summaries, are available in the supplementary materials (see Appendix Section B).

3.2 Model Justification

Each predictor in the model represents a specific condition under which a collision occurred, with hypothesized effects on injury severity. For driver condition, we expect that "Impaired" and "Distracted/Impaired" drivers are more likely to cause severe or fatal collisions due to reduced driving ability. Similarly, drivers under "Other" conditions may face various distractions or impairments, potentially increasing the severity of collisions. Conversely, drivers classified as "Normal" are anticipated to exhibit a reduced risk of severe outcomes.

Regarding road surface condition, we hypothesize that hazardous surfaces, such as ice or snow, may reduce injury severity because drivers often reduce speed in these conditions. Wet surfaces, however, may slightly increase severity due to decreased traction, although this effect may be less pronounced than hazardous conditions.

For light condition, we hypothesize that poor visibility under artificial or low lighting conditions may elevate injury severity, while natural light may reduce severity by improving visibility.

4 Results

Our results are summarized in Table 3.

4.0.1 Results

The results of the logistic regression model, as shown in Table 3, illustrate the relationships between injury severity in motor vehicle collisions and the predictor variables: driver condition, road surface condition, and light condition.

The **intercept** of the model is -0.820, which represents the baseline log-odds of a severe injury occurring when all other predictors are at their reference levels. The coefficients for each predictor indicate how the log-odds of a severe injury change with respect to the baseline.

4.0.1.1 Driver Condition

- **Impaired**: The coefficient for `DRIVCOND_GROUPImpaired` is 0.946 (SE = 0.306), suggesting that impaired driving significantly increases the likelihood of severe injuries in collisions compared to the baseline category (likely "Normal"). This positive coefficient indicates that impaired driving has a considerable association with severe outcomes.
- **Normal**: The coefficient for `DRIVCOND_GROUPNormal` is -0.730 (SE = 0.207), suggesting a decreased likelihood of severe injury relative to other conditions.
- **Other**: With a coefficient of 1.352 (SE = 0.169), "Other" driver conditions are also associated with a higher risk of severe injury, even more so than impaired driving.

Table 3: Explanatory model Injury Severity Prediction (n = 1000)

	Injury Severity
(Intercept)	−0.820
	(0.446)
DRIVCOND_GROUPImpaired	0.946
	(0.306)
DRIVCOND_GROUPNormal	−0.730
	(0.207)
DRIVCOND_GROUPOther	1.352
	(0.169)
RDSFCOND_GROUPHazardous Surface	−3.282
	(0.489)
RDSFCOND_GROUPOther	−2.864
	(0.410)
RDSFCOND_GROUPWet	−2.898
	(0.418)
LIGHT_GROUPNatural Light	−0.186
	(0.081)
LIGHT_GROUPOther	−0.043
	(0.098)
Num.Obs.	18 957
R2	0.029
Log.Lik.	−3570.692
ELPD	−3580.0
ELPD s.e.	83.1
LOOIC	7160.1
LOOIC s.e.	166.2
WAIC	7160.0
RMSE	0.22

4.0.1.2 Road Surface Condition

- **Hazardous Surface**: The coefficient for `RDSFCOND_GROUPHazardous Surface` is -3.282 (SE = 0.489), indicating a significant decrease in the likelihood of severe injury compared to "Good Condition" surfaces. This result may suggest that drivers exercise more caution or reduce speed on hazardous surfaces, potentially mitigating the severity of collisions.
- **Other**: The coefficient for `RDSFCOND_GROUPOther` is -2.864 (SE = 0.410), which also shows a reduction in the likelihood of severe injury under unspecified surface conditions.
- **Wet**: The coefficient for `RDSFCOND_GROUPWet` is -2.898 (SE = 0.418), similarly indicating a reduced risk of severe injury on wet surfaces.

4.0.1.3 Light Condition

- **Natural Light**: The coefficient for `LIGHT_GROUPNatural Light` is -0.186 (SE = 0.081), suggesting a slight reduction in the likelihood of severe injury compared to artificial light, though this effect is smaller than for the road surface conditions.
- **Other**: With a coefficient of -0.043 (SE = 0.098), the "Other" category for light conditions does not show a significant impact on injury severity, suggesting that lighting variations may have minimal influence on collision severity in this dataset.

Overall, the model results provide insights into how different conditions are associated with the severity of injuries in motor vehicle collisions. Factors such as impaired driving and unspecified road conditions significantly increase the likelihood of severe outcomes, while hazardous and wet surfaces appear to be associated with a reduced risk of severe injuries, possibly due to changes in driver behavior under these conditions.

5 Discussion

5.1 Relationship between Driver Condition and Injury Severity

The analysis shows that driver condition plays a crucial role in determining the severity of injuries in motor vehicle collisions. Drivers categorized as "Impaired" or "Other" are associated with a significantly higher risk of severe injuries, supporting existing research on the dangers of impaired driving. Studies have shown that alcohol and drug impairment reduce reaction times and decision-making abilities, greatly increasing the likelihood of serious collisions (**NHTSA_2021?**). In Toronto, where this data was collected, impaired driving remains one of the primary causes of road fatalities, underscoring the importance of targeted interventions to address this issue. Policies such as stricter DUI (Driving Under the Influence) regulations, public awareness campaigns, and enhanced enforcement could help mitigate the risks associated with impaired driving. Conversely, "Normal" driving conditions show a lower likelihood of severe injury, emphasizing the impact of driver attentiveness and sobriety on road safety.

5.2 Relationship between Road Surface Condition and Injury Severity

Road surface conditions also influence injury severity in collisions, although perhaps not in the ways one might expect. The results indicate that hazardous surfaces, like icy or snowy roads, are associated with a lower likelihood of severe injuries. This finding could be explained by the fact that drivers may adjust their behavior on hazardous surfaces by reducing speed or increasing caution. Research suggests that adverse road conditions encourage safer driving practices, which can counterintuitively reduce injury severity in the event of an accident (**RoadSafety?**). On the other hand, wet surfaces still increase the likelihood of severe injuries compared to good conditions, likely due to reduced tire traction leading to loss of control. These insights suggest that public safety campaigns encouraging cautious driving during adverse weather conditions could be effective, alongside infrastructure improvements such as better drainage to reduce wet road hazards.

5.3 Relationship between Light Condition and Injury Severity

The model results indicate that light conditions play a minor role in injury severity, with a slight reduction in risk associated with natural light. This finding aligns with existing research showing that visibility is generally better in natural light, allowing drivers to detect and respond to hazards more effectively. Artificial light and "Other" lighting conditions are associated with a marginal increase in injury severity, which may be attributed to reduced visibility in low-light conditions or under artificial lighting (**LightingSafety?**). While the effect is smaller than for driver and road conditions, this finding highlights the importance of well-maintained and properly illuminated roadways, especially in high-traffic areas and intersections.

5.4 Limitations

5.4.1 Data Limitations Due to Geographic Offsetting

One limitation of this dataset is the geographic offsetting of incident locations to protect privacy. While necessary for privacy concerns, this adjustment may reduce the accuracy of analyses that focus on specific intersections or neighborhoods. For instance, analyses attempting to identify collision "hotspots" or evaluate the effectiveness of localized safety measures could be skewed. Future research may benefit from accessing more precise location data where privacy can still be safeguarded through other methods, such as anonymization.

5.4.2 Underrepresentation of Non-Severe Incidents

The dataset only includes collisions that resulted in severe injuries or fatalities, excluding minor incidents that also provide valuable insights into road safety. This focus on high-severity incidents may lead to an overrepresentation of conditions associated with severe outcomes, potentially skewing the findings. Incorporating data on less severe incidents would allow for a more comprehensive analysis of risk factors across all collision types, enhancing the ability to develop preventive measures that address both severe and minor incidents.

5.5 Future Steps

Further research should consider expanding the dataset to include a broader range of collision outcomes, capturing minor injuries and non-injury collisions. This would provide a more holistic view of road safety in Toronto, identifying conditions that contribute to all types of collisions. Additionally, exploring interactions between environmental and human factors—such as how road surface conditions might interact with driver behavior—could yield more nuanced insights. Policymakers and city planners could benefit from such research to implement multifaceted safety measures, ultimately creating a safer driving environment in Toronto.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4 we implement a posterior predictive check. This shows...

In Figure 5 we compare the posterior with the prior. This shows...

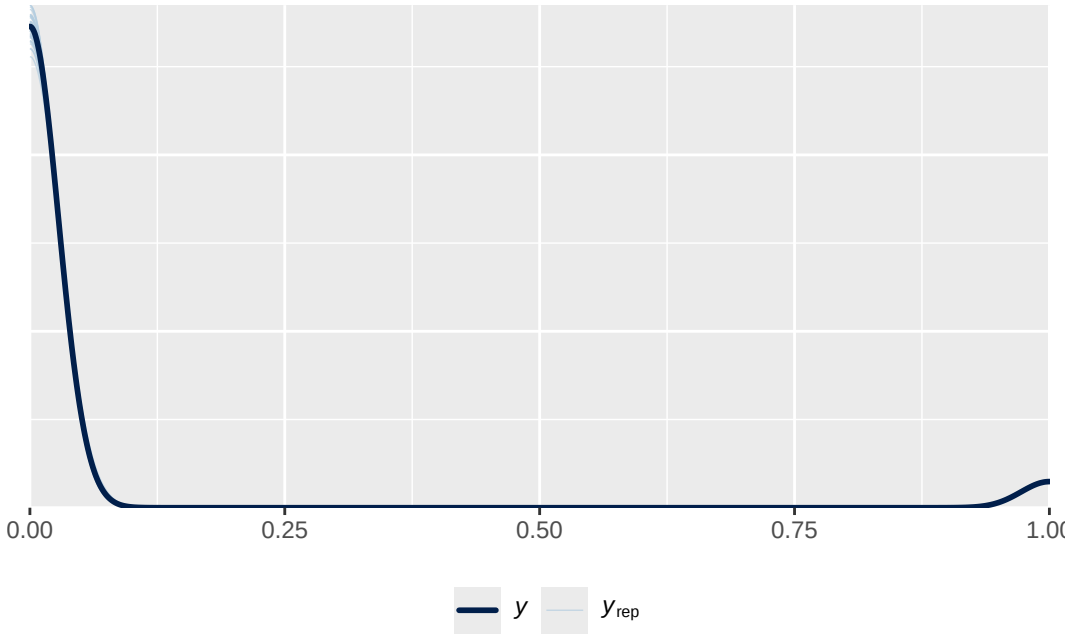


Figure 4: Posterior distribution for logistic regression model

B.2 Diagnostics

Figure 6 and Figure 7 are trace plots. It shows... This suggests...

Figure 8 is a Rhat plot. It shows... This suggests...

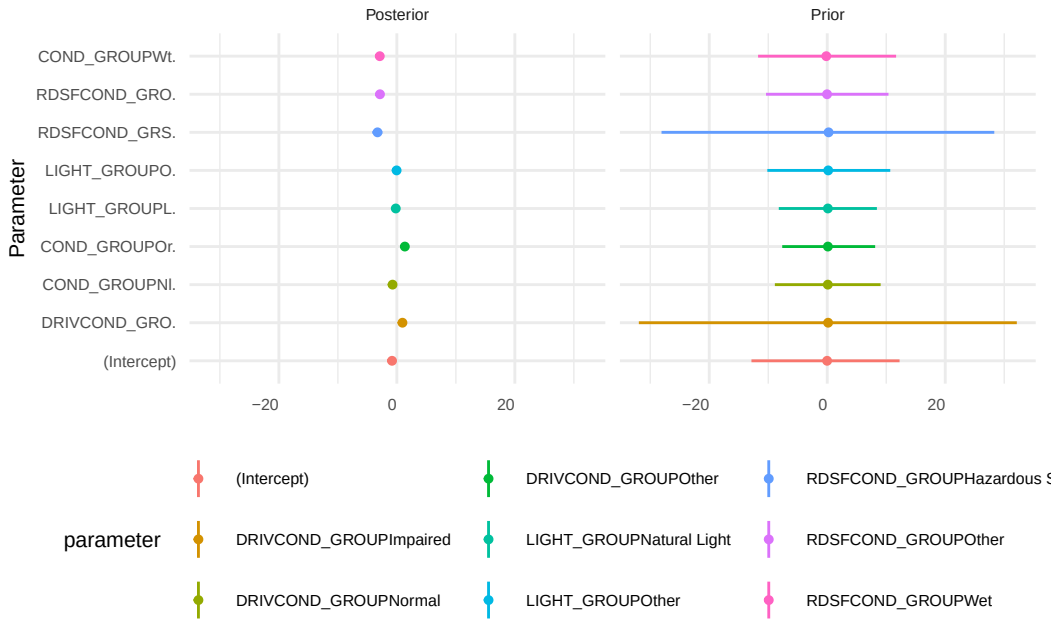


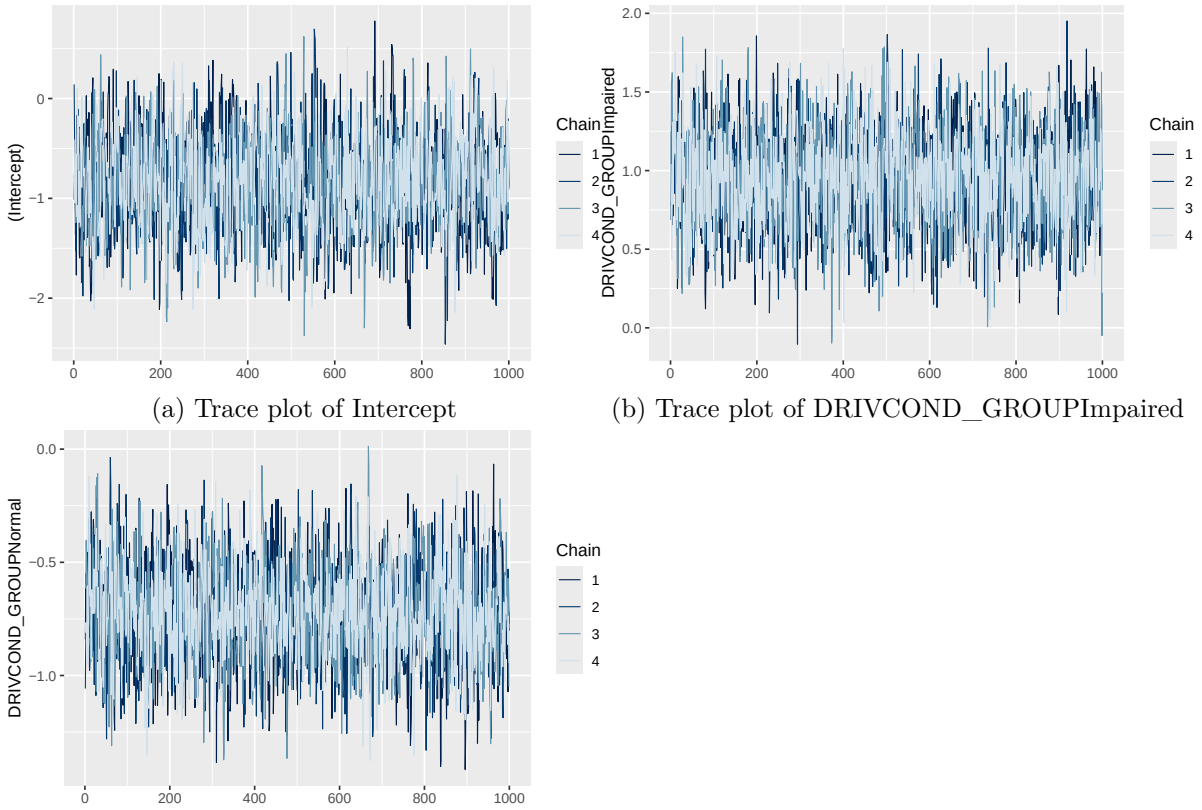
Figure 5: Comparing the posterior with the prior

B.3 Markov Chain Monte Carlo Convergence Check

C Trace plots and Rhat plots for the `injury_severity_model`

C.1 90% Credibility Interval

Figure 9 and Figure 10 are 90% credibility interval plots for the injury severity predictors.

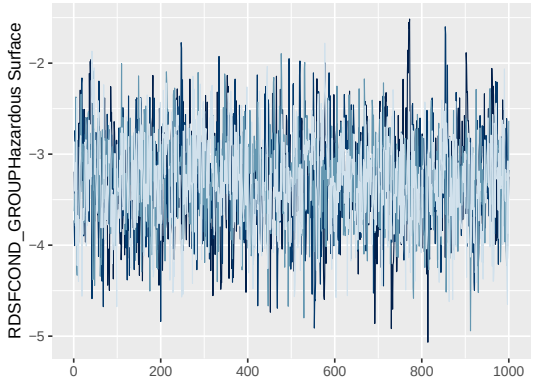


(a) Trace plot of Intercept

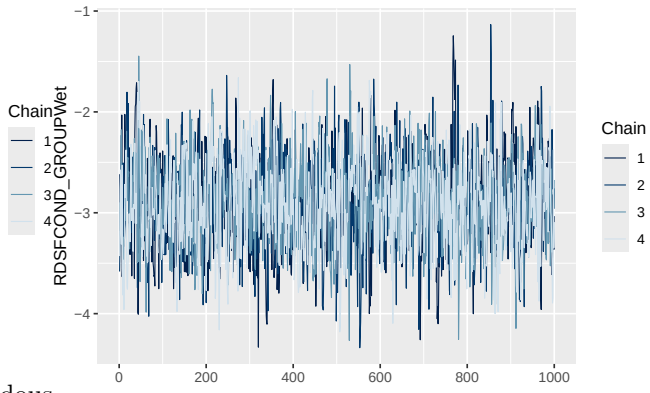
(b) Trace plot of DRIVCOND_GROUPImpaired

(c) Trace plot of DRIVCOND_GROUPNormal

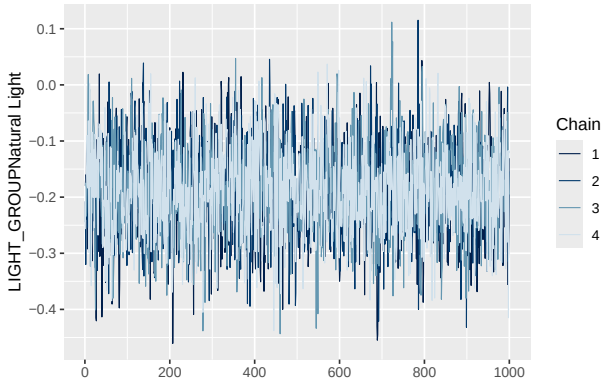
Figure 6: Trace plot of intercept and driver condition



(a) Trace plot of RDSFCOND_GROUPHazardous Surface



(b) Trace plot of RDSFCOND_GROUPWet



(c) Trace plot of LIGHT_GROUPNatural Light

Figure 7: Trace plot of road surface and light condition

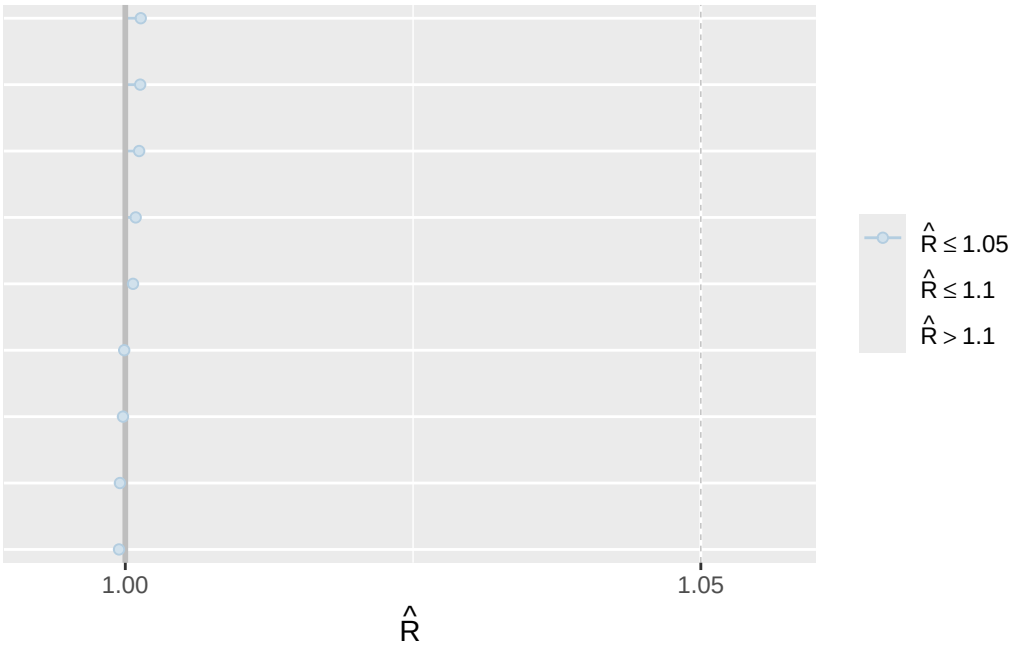


Figure 8: Rhat plot for MCMC convergence

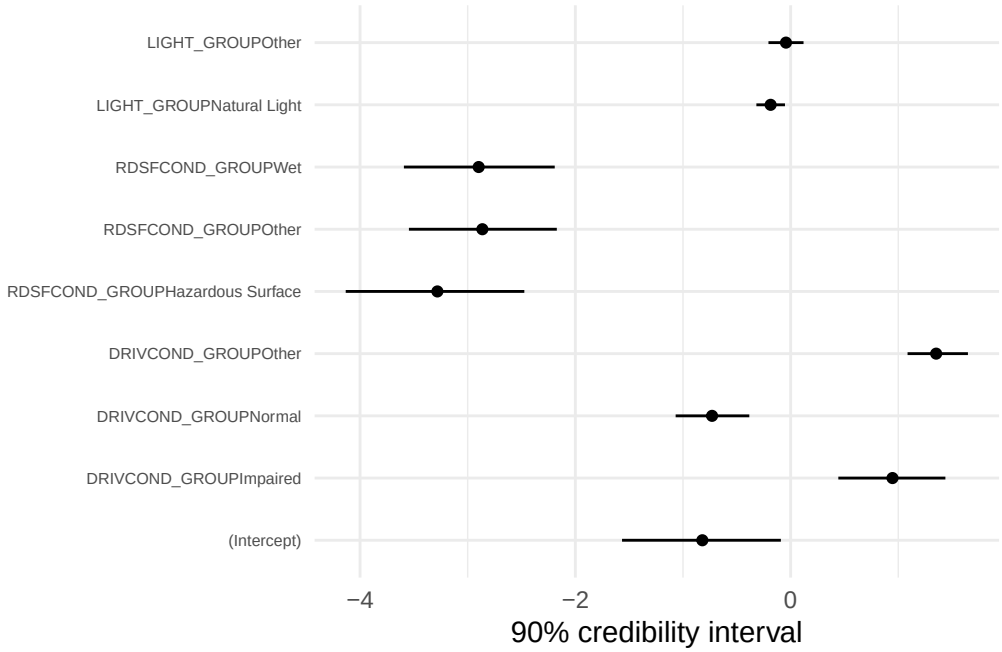


Figure 9: Credible intervals for predictors of injury severity

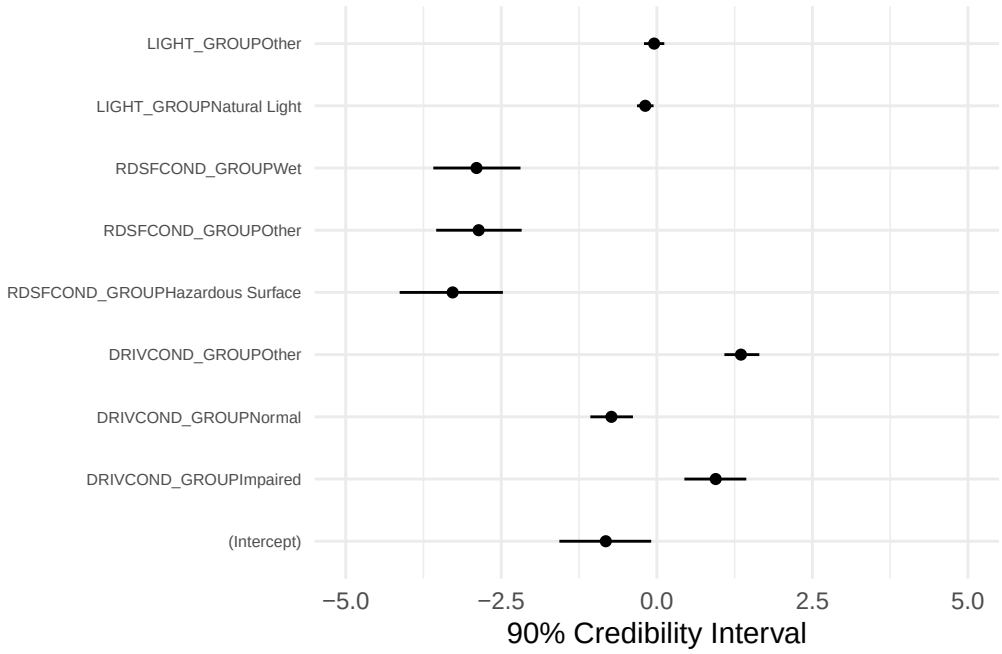


Figure 10: Credible intervals for predictors of injury severity with x-axis limits

References

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

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.