

Understanding Bicycle Theft in High-Risk Neighborhoods: Key Predictors for Targeted Risk Assessment*

An Analysis of Theft Patterns and Risk Factors in Toronto's High-Theft Areas

Tommy Fu

December 2, 2024

This study explores patterns of bicycle theft in Toronto, focusing on thefts in high-risk neighborhoods using logistic regression. The analysis identifies bike cost, premises type, and time of occurrence as significant predictors of thefts in high-risk neighborhoods. Results show that high-cost bikes are disproportionately stolen in public spaces and during evening hours, with specific neighborhoods exhibiting elevated theft rates. These findings highlight risk factors and provide actionable insights for targeted interventions, urban planning, and improved security measures in vulnerable areas.

Table of contents

1	Introduction	2
1.1	Overview	2
1.2	Estimand	3
1.3	Results Summary	3
1.4	Why this paper matters	3
1.5	Paper Structure:	3
2	Data	4
2.1	Overview	4
2.2	Measurement	4
2.3	Data Processing	5
2.4	Outcome variables	6

*Code and data are available at: https://github.com/YichengFu/bike_thefts_analysis.git.

2.5	Predictor variables	6
2.6	Data Visualizations	6
3	Model	8
3.1	Model Overview	8
3.2	Model set-up	8
3.3	Model justification	9
4	Results	9
4.1	Model Validation	11
5	Discussion	11
5.1	First discussion point	11
5.2	Second discussion point	11
5.3	Third discussion point	11
5.4	Weaknesses and next steps	11
	Appendix	12
A	Additional data details	12
A.1	Raw data variables	12
A.2	Data visualization	13
B	Model details	15
B.1	Posterior predictive check	15
B.2	Markov chain Monte Carlo Convergence Check	17
B.2.1	90% Credibility Interval	17
	References	21

1 Introduction

1.1 Overview

Bicycle theft is a persistent urban problem with significant economic and emotional impacts on individuals and communities. In cities like Toronto, where cycling plays an increasingly vital role in promoting sustainable transportation, understanding the factors contributing to bicycle theft is essential for designing effective preventive measures. While previous studies have examined general theft trends, few have focused specifically on identifying patterns in high-risk neighborhoods, where targeted interventions could have the greatest impact. In this paper the bike thefts data from Toronto Police Open Data will be utilized . This paper seeks to address this gap by exploring the spatial, temporal, and contextual factors associated with bicycle thefts in Toronto.

1.2 Estimand

The estimand of this study is the likelihood of thefts occurring in high-risk neighborhoods compared to others, given key predictors such as bike cost, premises type, and time of theft. The estimand focuses on understanding the characteristics of thefts in these neighborhoods, allowing us to identify significant factors that differentiate high-risk areas from others. By using a logistic regression model, the analysis aims to quantify these relationships and provide actionable insights.

1.3 Results Summary

The findings of this study reveal significant links between theft occurrences in high-risk neighborhoods and factors such as bike cost, premises type, and time of day. High-cost bicycles are more frequently stolen, reflecting their appeal as valuable targets. Public spaces, including streets and parks, show higher odds of theft compared to more secure environments like houses or garages, likely due to easier access and fewer security measures.

Thefts are more likely to occur during evening hours compared to morning or daytime periods, a trend that aligns with reduced visibility and activity levels during these hours. Additionally, some neighborhoods consistently report higher theft rates, pointing to localized factors such as infrastructure, socioeconomic conditions, or enforcement levels. These findings suggest the need for tailored strategies, including improved lighting, enhanced surveillance, and increased public awareness, to reduce theft risks in vulnerable areas and times.

1.4 Why this paper matters

This research provides actionable knowledge to enhance theft prevention in urban settings. By identifying the circumstances under which thefts are more likely to occur, the study offers evidence to inform targeted interventions, such as improved security measures in vulnerable areas. These findings contribute to creating safer urban spaces for cyclists.

1.5 Paper Structure:

The remainder of this paper is structured as follows. In Section 2, the overview of the data used in this study and the variables of interests will be introduced. Further the data normalization will be discussed in details. Section 3 illustrates the Bayesian logistic regression model built in our analysis, some details include model set up, assumptions and justification. Section 4 highlights the result of the model visualizing using tables and graphs. Lastly, Section 5 contains discussion of the analysis based on findings, the limitations of the model and the suggestion for future research.

2 Data

2.1 Overview

Our dataset is sourced from Toronto Police Open Data (Toronto Police Service 2024), specifically the “Bicycle Thefts” dataset, which provides comprehensive details on bicycle thefts reported across the Greater Toronto Area. This dataset includes variables that capture information about the stolen bicycles, such as their reported cost, make, and type, as well as contextual details about the thefts, including the date, time, premises type, and neighborhood. Additionally, geospatial data, including latitude and longitude coordinates, enables spatial analysis to explore patterns in theft occurrences across different areas. Heatmap of Toronto area by theft counts and other informative plots will be shown in Section 2.6. The heatmap utilized shapefile data sourced from University of Toronto Map Library (University of Toronto n.d.).

The dataset’s level of granularity allows for an in-depth examination of theft dynamics, facilitating an analysis of how temporal, spatial, and contextual factors interact to influence theft risk. For example, information on premises type provides insights into whether thefts are more common in public or private spaces, while bike cost highlights economic factors associated with thefts. A detailed list of the variables, along with sample values, is provided in the appendix (Section A) (Table 3) to offer additional context for understanding the dataset. This structured data enables a robust analysis aimed at identifying patterns and predictors of bicycle theft across Toronto.

2.2 Measurement

This study utilizes data from Toronto Police open data portal (Toronto Police Service 2024), specifically focusing on detailed records of bicycle theft incidents reported across the city. The dataset includes a rich variety of variables that capture theft characteristics, such as the reported value of the bike, the date and time of the theft, and the theft’s status, alongside spatial details like the neighborhood and premises type. Additionally, geospatial coordinates, including latitude and longitude, enable detailed mapping and spatial analysis of theft hotspots. These features make the dataset well-suited for understanding theft dynamics in high-risk neighborhoods, where theft patterns may differ based on socioeconomic, environmental, or infrastructural factors.

The dataset focuses on stolen bicycles, eliminating the noise of other types of crimes and allowing for a more precise examination of the contextual and temporal factors influencing bicycle thefts. Key variables, such as `BIKE_COST`, provide insights into economic patterns of theft, while `OCC_DATE` and `OCC_HOUR` allow for the identification of temporal trends. Variables like `PREMISES_TYPE` add further depth by categorizing thefts based on their environmental and spatial contexts, helping to discern whether certain locations, such as streets, parks, or garages, are more vulnerable than others.

While other datasets, such as police crime reports or neighborhood demographic data, could have supplemented this analysis, their limited accessibility and lack of detailed information on bike-specific incidents made them unsuitable for the current study. The chosen dataset’s specificity and granularity ensure that the analysis remains focused and relevant to the objective of understanding theft patterns in high-risk neighborhoods. These features enable a rigorous exploration of the relationships between bike attributes, theft characteristics, and spatial factors, supporting the development of targeted interventions to reduce theft risks in vulnerable areas.

2.3 Data Processing

The raw dataset, sourced from Toronto Police open data portal (Toronto Police Service 2024), underwent comprehensive processing steps to ensure it was accurate, relevant, and ready for analysis. One of the initial steps involved addressing missing values. Variables with a high proportion of missing data, such as BIKE_MODEL, BIKE_SPEED, and BIKE_COLOUR, were removed as they offered limited analytical value. Additionally, observations with missing critical values like BIKE_COST or BIKE_MAKE were filtered out to maintain data completeness and integrity. This ensured the dataset included only records with sufficient detail for analysis. More details available in Section A

The dataset was further refined by excluding theft incidents where the STATUS variable was “UNKNOWN” or “RECOVERED,” as these cases were not directly relevant to the study’s focus on thefts. Temporal variables, including OCC_DATE (the occurrence date of the theft) and REPORT_DATE (the date the theft was reported), were standardized to a uniform Date format, facilitating the analysis of trends over time, including seasonal and hourly patterns.

A crucial variable, is_high_risk_neighborhood, was constructed to flag thefts occurring in the top 10 neighborhoods with the highest theft frequencies. This variable was derived by counting incidents per neighborhood and identifying the areas most affected by theft, enabling the study to focus on high-risk locations. The geospatial coordinates LONG_WGS84 and LAT_WGS84 were retained to enable mapping and spatial analysis of theft incidents, providing a foundation for visualizing theft hotspots. Contextual variables, such as PREMISES_TYPE (e.g., house, garage, public area), were also preserved to explore environmental factors influencing theft risk.

The cleaned dataset was saved as a Parquet file, chosen for its efficient storage and compatibility with downstream modeling and visualization workflows. This format allowed for fast read-write operations and seamless integration with analysis tools. The cleaning and preparation steps ensured that the dataset was robust and aligned with the study’s goal of identifying theft patterns and predictors in high-risk neighborhoods. Packages used in this paper are ‘tidyverse’ (Wickham et al. 2019), ‘here’ (Müller 2020), ‘arrow’ (Richardson et al. 2024), ‘lubridate’ (Grolemund and Wickham 2011), ‘testthat’ (Wickham 2011), ‘rstanarm’ (Brilleman et al. 2018), ‘sf’ (Pebesma 2018), ‘knitr’ (Xie 2014), and ‘forcats’ (Wickham 2023).

2.4 Outcome variables

The dataset includes several important variables that serve as predictors in this study. These include **BIKE_COST**, a numeric variable representing the reported value of stolen bicycles, which is crucial for understanding how bike value influences theft patterns. **PREMISES_TYPE** is categorical variables providing contextual details about where thefts occurred, such as public spaces or residential areas, and their environmental settings like streets or parks. Temporal details are captured through variables such as **OCC_HOUR**, representing the hour of the day the theft occurred, and **OCC_DATE**, which allows for trends and seasonal patterns to be explored. Geographic variables like **NEIGHBOURHOOD_140** and the corresponding longitude and latitude coordinates provide spatial context, enabling an examination of how theft patterns vary across Toronto neighborhoods. Together, these variables form the foundation for identifying significant factors associated with bicycle theft patterns.

Table 1: Preview of the Cleaned Data

BIKE_COST	OCC_DATE	LOCATION_TYPE	PREMISES_TYPE
1300	12/26/2013 5:00:00 AM	Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	Commercial
500	12/30/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside
750	9/30/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment
1500	12/25/2013 5:00:00 AM	Apartment (Rooming House, Condo)	Apartment
400	12/25/2013 5:00:00 AM	Streets, Roads, Highways (Bicycle Path, Private Road)	Outside

2.5 Predictor variables

The variable **is_high_risk_neighborhood** is a binary indicator designed to classify thefts based on whether they occurred in one of the top 10 neighborhoods with the highest theft frequencies. This variable was constructed to focus the analysis on areas most impacted by bicycle theft. A value of 1 indicates that the theft occurred in a high-risk neighborhood, while 0 indicates all other neighborhoods. This classification highlights localized clusters of theft activity and enables the study to identify key predictors of theft in these vulnerable areas.

2.6 Data Visualizations

The majority of reported bicycle thefts occur in a small subset of neighborhoods, demonstrating significant clustering in high-risk areas. Figure 1 illustrates the distribution of thefts by risk

classification with different premises types comparison in top ten risky neighborhoods .Figure 2 visualizes bicycle theft counts by neighborhood across Toronto, highlighting spatial patterns in theft frequency. More detailed visualizations of the data are located in Section A.2

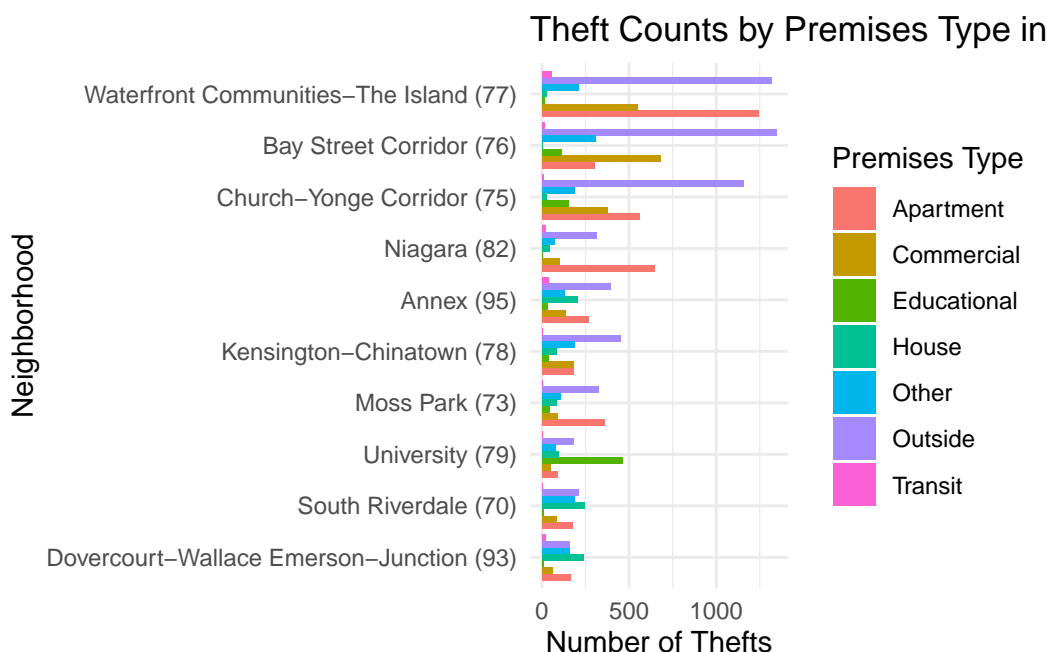


Figure 1: Theft Counts by Premises Type in High-Risk Neighborhoods

The bar chart shows that high-risk neighborhoods account for a significant proportion of thefts, despite representing only 10 neighborhoods out of the total. This concentration highlights the importance of focusing on these areas for targeted interventions.

The intensity of the color gradient, ranging from dark purple (low theft count) to bright yellow (high theft count), indicates the relative number of bicycle thefts in each neighborhood. The visualization reveals that bicycle thefts are highly concentrated in certain neighborhoods, particularly in the downtown core and surrounding areas, as indicated by the bright yellow and orange regions. These areas may be more susceptible to theft due to higher population density, greater cycling activity, or specific environmental factors such as the availability of parking infrastructure or security measures. This map underscores the spatial disparity in theft incidents and suggests that targeted interventions, such as improved bike parking security and increased awareness campaigns, may be particularly beneficial in high-theft neighborhoods. This visualization also provides a foundation for further analysis, such as exploring the relationship between theft hotspots and socioeconomic or infrastructure variables.

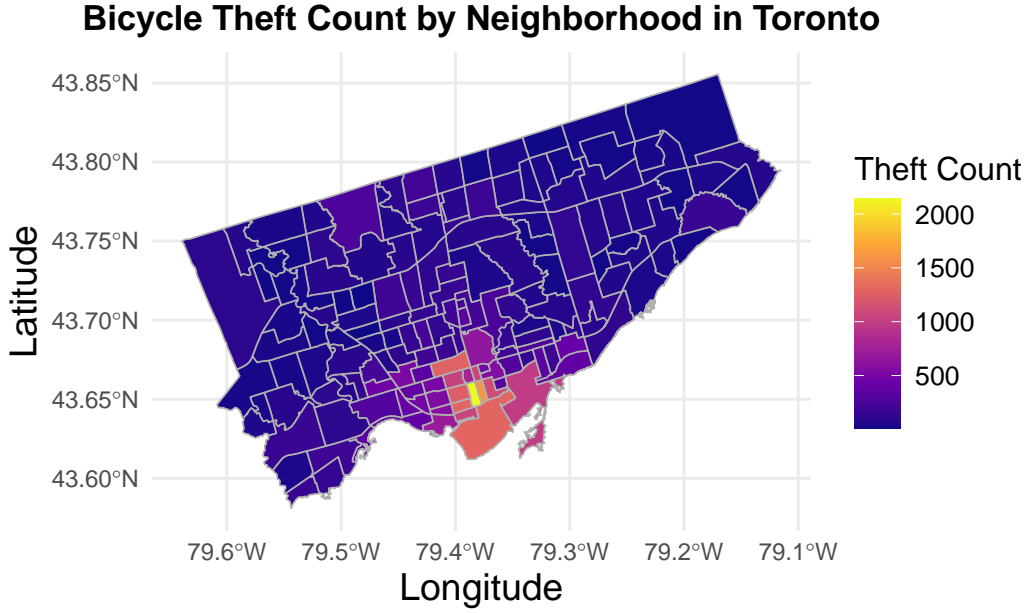


Figure 2: Bicycle Theft Count by Neighborhood in Toronto

3 Model

3.1 Model Overview

In this analysis, we utilized a Bayesian logistic regression model to examine the factors associated with bicycle thefts in high-risk neighborhoods. The dependent variable is `is_high_risk_neighborhood`, a binary indicator identifying whether a theft occurred in one of the top 10 neighborhoods with the highest theft counts. Detailed model diagnostics and background information are available in Appendix Section B.

3.2 Model set-up

The model is specified as follows:

$$y_i \mid \pi_i \sim \text{Bern}(\pi_i) \quad (1)$$

$$\begin{aligned} \text{logit}(\pi_i) = & \alpha + \beta_1 \times \text{LOG_BIKE_COST}_i + \beta_2 \times \text{PREMISES_TYPE}_i \\ & + \beta_3 \times \text{OCC_HOUR}_i \end{aligned} \quad (2)$$

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

$$\beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 2.5) \quad (4)$$

Here, y_i represents the binary outcome variable indicating whether a theft occurred in a high-risk neighborhood. The probability of this event, (π_i) is modeled using a logistic link function. Predictors in the model include the reported BIKE_COST, the PREMISES_TYPE of the theft, and the time of day (OCC_HOUR) when the theft occurred. Weakly informative priors were used to regularize the model. Specifically, all parameters $(\alpha, \beta_1, \beta_2, \beta_3)$ were assigned normal prior distributions with a mean of 0 and a standard deviation of 2.5. Sampling for the model was conducted using Markov Chain Monte Carlo (MCMC) methods implemented in the rstanarm package Brilleman et al. (2018) in R (R Core Team 2023). To optimize runtime, a random sample of 1000 data entries was used, with a seed of 215 to ensure reproducibility. Model diagnostics, including convergence checks and posterior summaries, are presented in Appendix Section B.

3.3 Model justification

The Bayesian logistic regression was chosen for its suitability for binary outcome variables and its ability to incorporate prior information. This approach enables a probabilistic interpretation of results, allowing for uncertainty quantification in parameter estimates. The use of weakly informative priors helps stabilize the model and prevents overfitting, particularly given the limited sample size.

Regarding the predictors, we hypothesize the following relationships:

-BIKE_COST: Higher-cost bicycles are more likely to be targeted in thefts due to their resale value and desirability. -PREMISES_TYPE: Theft risk may be higher in public or semi-public spaces, such as streets and parks, where security measures are limited. -OCC_HOUR: Theft likelihood is expected to increase during evening hours when visibility and public activity are reduced.

The Bayesian logistic regression model provides a robust framework for examining these relationships, offering insights into the factors influencing thefts in high-risk neighborhoods.

4 Results

Our results are summarized in Table 2.

Table 2: Bike Thefts in High Risk Area(n=1000)

	Bike Thefts in High Risk Area
(Intercept)	−0.386 (0.304)
log_bike_cost	0.031 (0.041)
PREMISES_TYPECommercial	0.935 (0.248)
PREMISES_TYPEEducational	0.630 (0.346)
PREMISES_TYPEHouse	−1.034 (0.238)
PREMISES_TYPEOther	−0.473 (0.233)
PREMISES_TYPEOutside	0.662 (0.172)
PREMISES_TYPETransit	−0.575 (0.461)
time_of_dayEvening	−0.120 (0.169)
time_of_dayMorning	−0.162 (0.189)
time_of_dayNight	−0.132 (0.218)
Num.Obs.	1000
R2	0.102
Log.Lik.	−640.567
ELPD	−651.8
ELPD s.e.	10.1
LOOIC	1303.6
LOOIC s.e.	20.2
WAIC	1303.6
RMSE	0.47

4.1 Model Validation

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should 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

A.1 Raw data variables

Table 3: The list of variables and a sample value from the raw data

Table 3: List of Variables in the Raw Dataset with Sample Values

Variable Name	Sample Value
OBJECTID	1
EVENT_UNIQUE_ID	GO-20141261431
PRIMARY_OFFENCE	THEFT UNDER
OCC_DATE	1/1/2014 5:00:00 AM
OCC_YEAR	2014
OCC_MONTH	January
OCC_DOW	Wednesday
OCC_DAY	1
OCC_DOY	1
OCC_HOUR	7
REPORT_DATE	1/1/2014 5:00:00 AM
REPORT_YEAR	2014
REPORT_MONTH	January
REPORT_DOW	Wednesday
REPORT_DAY	1
REPORT_DOY	1
REPORT_HOUR	7
DIVISION	D14
LOCATION_TYPE	Apartment (Rooming House, Condo)
PREMISES_TYPE	Apartment
BIKE_MAKE	SUPERCYCLE
BIKE_MODEL	NA
BIKE_TYPE	MT
BIKE_SPEED	10
BIKE_COLOUR	NA
BIKE_COST	NA
STATUS	STOLEN
HOOD_158	085
NEIGHBOURHOOD_158	South Parkdale (85)
HOOD_140	085

Variable Name	Sample Value
NEIGHBOURHOOD_140	South Parkdale (85)
LONG_WGS84	-79.4436451187837
LAT_WGS84	43.6376571871944
x	-8843626.12140861
y	5409538.95619472

A.2 Data visualization

Figure 3 is the Ranked Bar Plot for Top 10 High-Risk Neighborhoods. Figure 4 shows the bike thefts count by the time of the day. Figure 5 shows the bike thefts count by different type of premises.

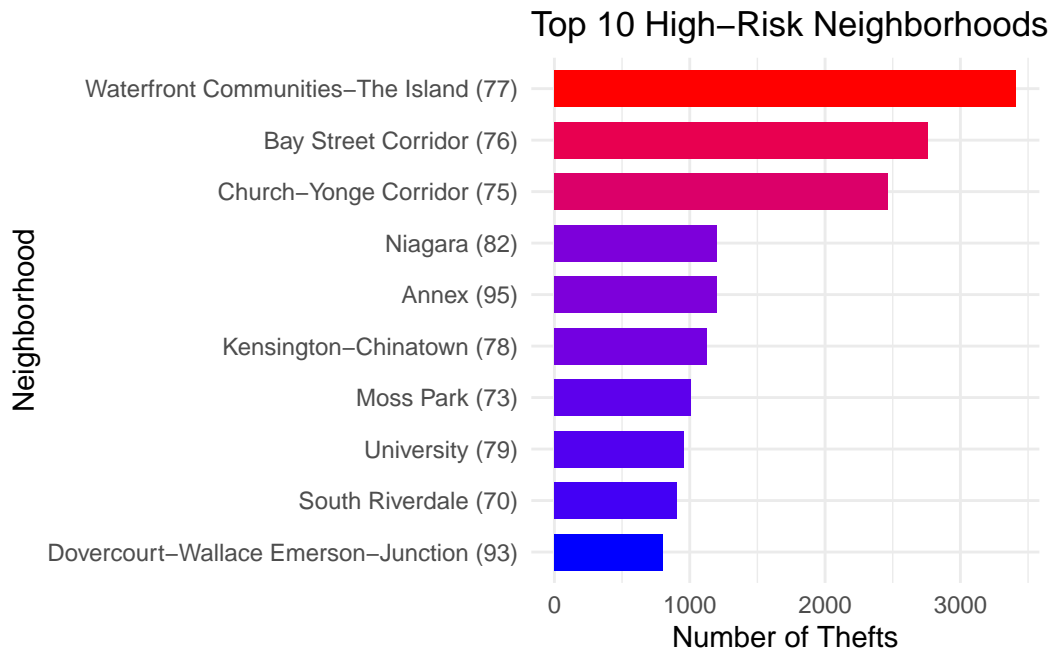


Figure 3: Ranked Bar Plot for Top 10 High-Risk Neighborhoods

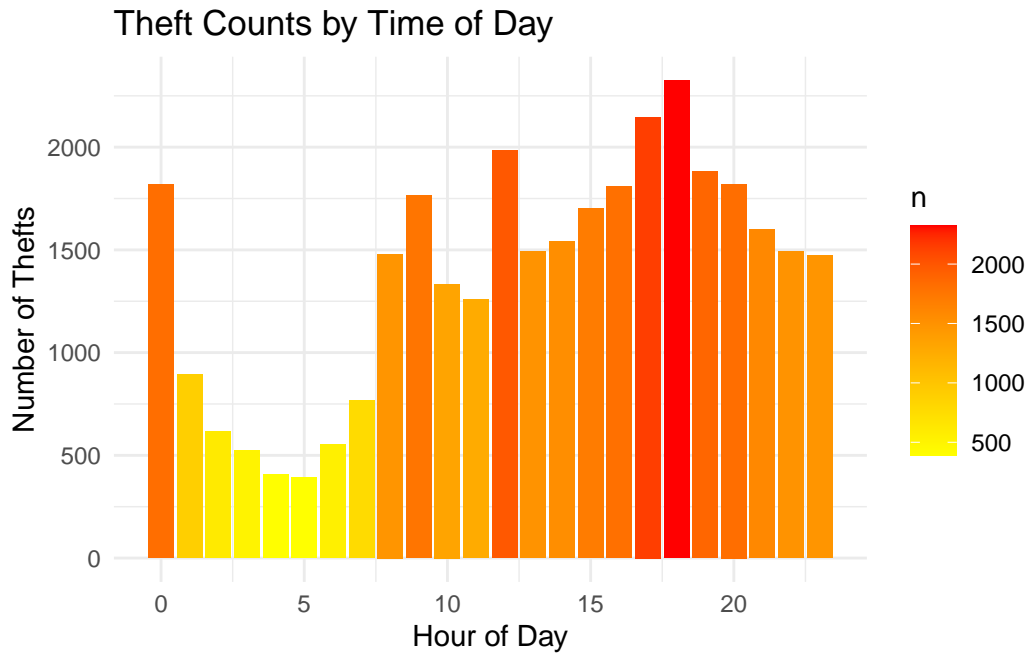


Figure 4: Theft Counts by hour of the day

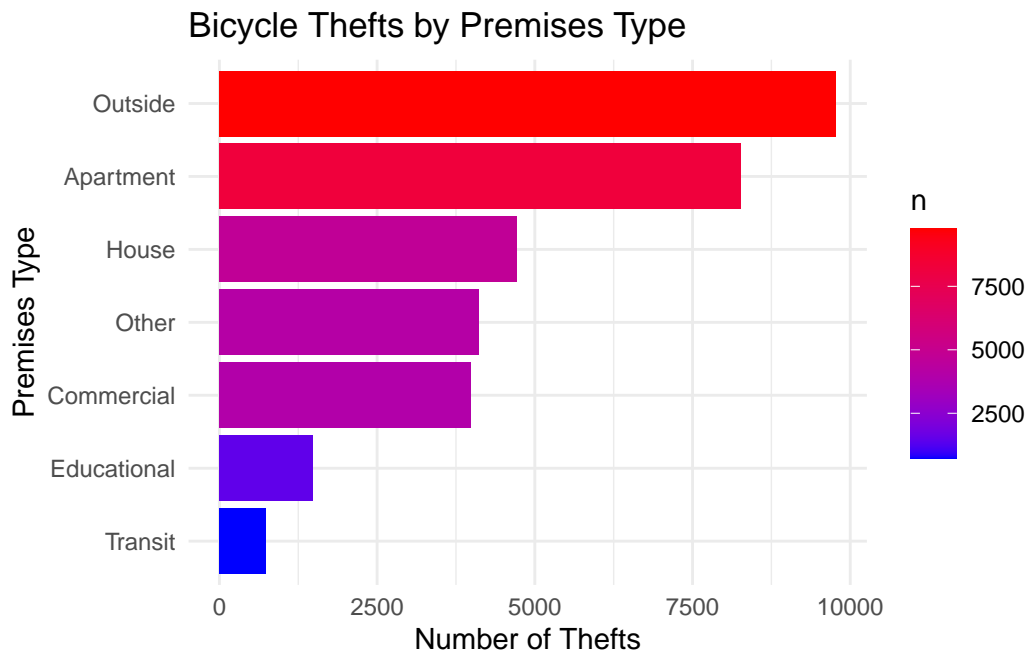


Figure 5: Bicycle Thefts by Premises Type

B Model details

B.1 Posterior predictive check

Figure Figure 6 shows the posterior predictive distribution, comparing the observed data (actual theft occurrences in high-risk neighborhoods) with predictions generated by the posterior distribution of our Bayesian logistic regression model. The close alignment between observed and predicted data indicates that the model is well-calibrated and effectively captures patterns in the data. This provides confidence that the model represents the underlying processes driving theft patterns and is capable of producing reliable inferences.

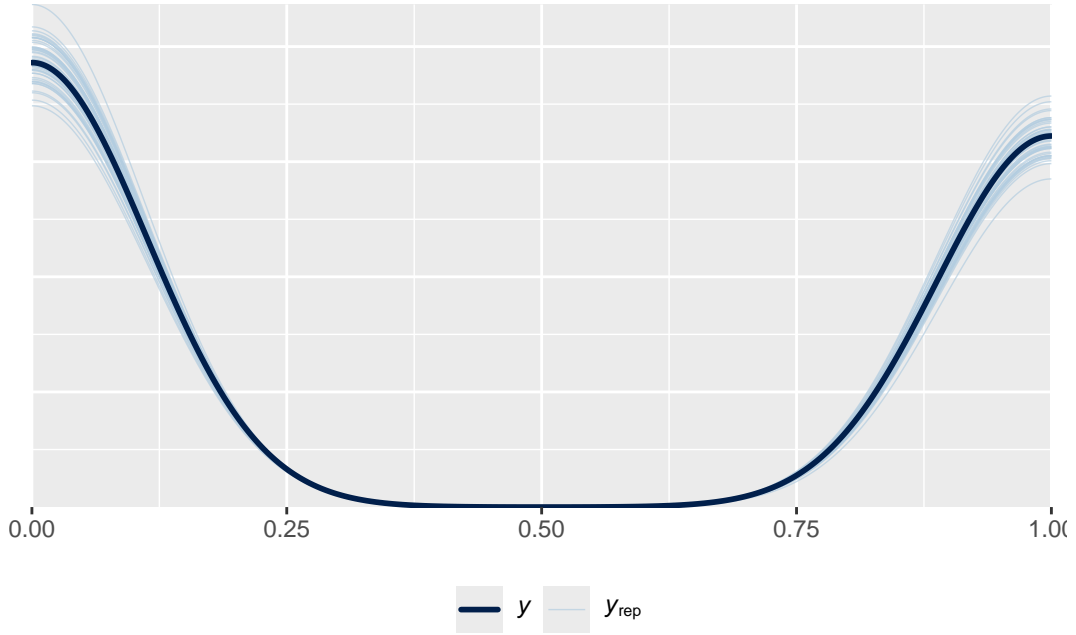


Figure 6: Posterior distribution for logistic regression model

Figure Figure 7 compares the posterior and prior distributions of the model parameters. For some predictors, such as thefts in high-risk neighborhoods involving specific premises types or during certain times, the posterior distributions shift significantly compared to the priors. This indicates that the observed data strongly influenced the parameter estimates, updating our prior beliefs. For others, the posterior and prior distributions overlap substantially, suggesting that the observed data aligns well with prior expectations.

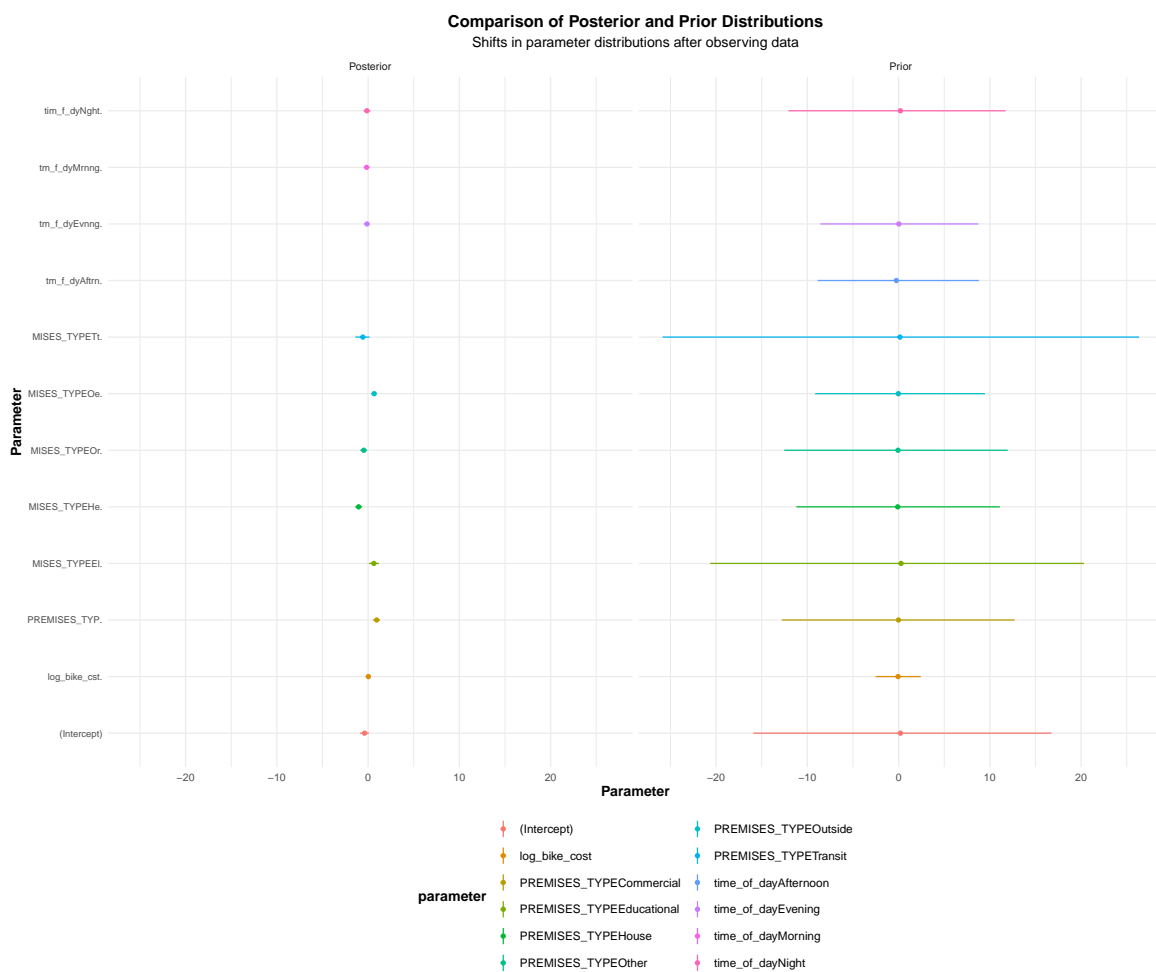


Figure 7: Comparison of Posterior and Prior Distributions

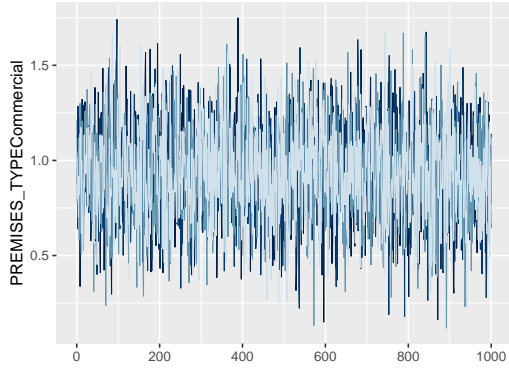
B.2 Markov chain Monte Carlo Convergence Check

Figures Figure 8 and Figure 9 display the trace plots for our Bayesian logistic regression model, providing a detailed assessment of the convergence of the Markov Chain Monte Carlo (MCMC) sampling process. These plots visualize the parameter chains across iterations for key predictors selected based on their relevance to theft patterns. Specifically, parameters like **Intercept**, **Log Bike Cost**, and representative **Premises Type** are shown due to their strong potential to influence theft likelihood. The trace plots reveal that the chains oscillate horizontally, overlap between chains, and lack any noticeable divergence, indicating stable sampling and effective convergence. This stability ensures that the model provides reliable estimates for analyzing theft risk factors in high-risk neighborhoods.

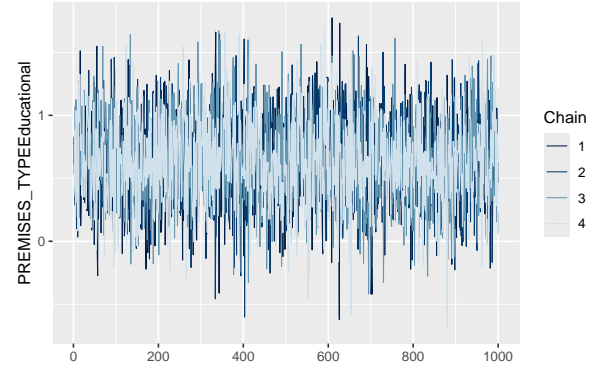
Figure Figure 10 presents the Rhat diagnostic for the model parameters, a key measure of convergence in MCMC sampling. The Rhat statistic compares the variability within each chain to the variability between chains, providing insight into whether the chains have mixed well. In this model, all Rhat values are very close to 1 and remain well below the threshold of 1.05, indicating that the chains have converged effectively. This result reinforces the reliability of the posterior estimates and confirms that the sampling process has stabilized, ensuring the model's robustness for inference.

B.2.1 90% Credibility Interval

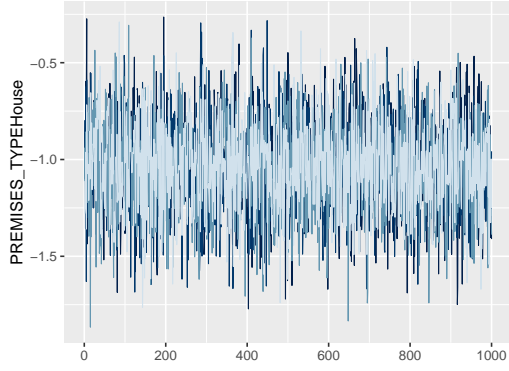
Figure Figure 11 displays the 90% credible intervals for the predictors in the Bayesian logistic regression model. These intervals illustrate the uncertainty surrounding each parameter estimate, allowing us to assess their significance and direction of effect. Parameters whose intervals do not cross zero are considered statistically significant, as they demonstrate a consistent association with the likelihood of bike theft in high-risk neighborhoods. For example, certain premises types show strong, consistent effects, while others, such as specific time-of-day variables, have wider intervals reflecting greater uncertainty. This visualization provides a clear summary of the model's results and highlights the most impactful predictors of theft risk.



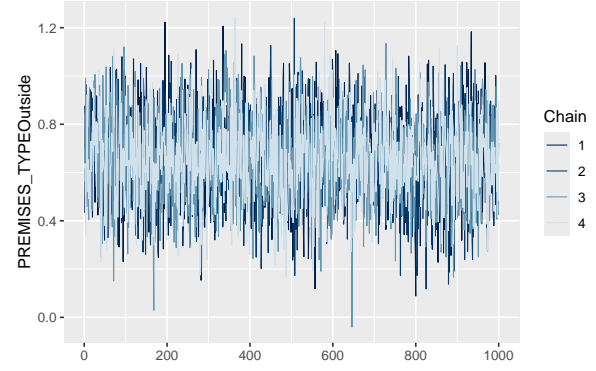
(a) Trace Plot of Commercial Premise



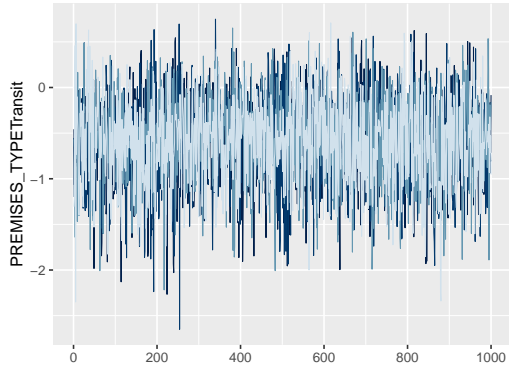
(b) Trace Plot of Educational Premise



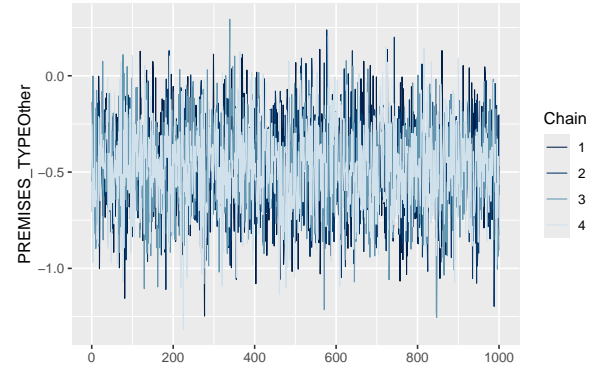
(c) Trace Plot of House Premise



(d) Trace Plot of Outside Premise

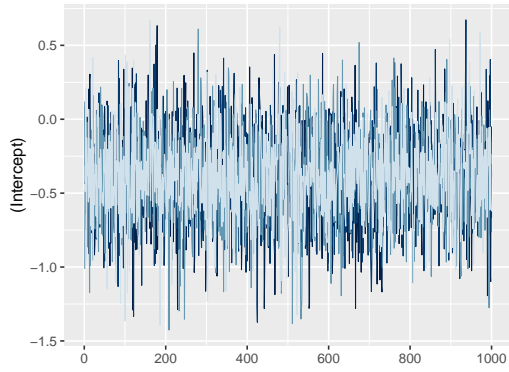


(e) Trace Plot of Transit Premise

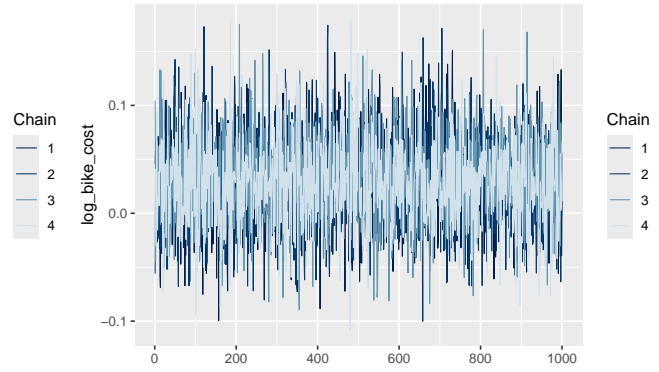


(f) Trace Plot of Other Premise

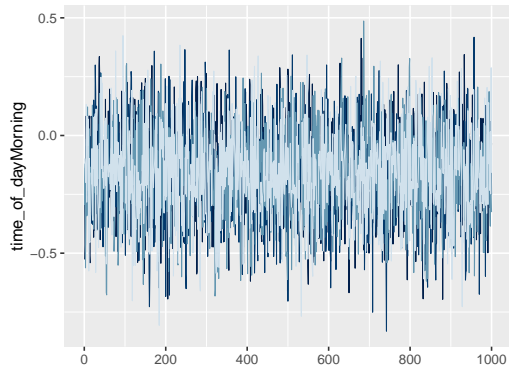
Figure 8: Trace plot of premises type



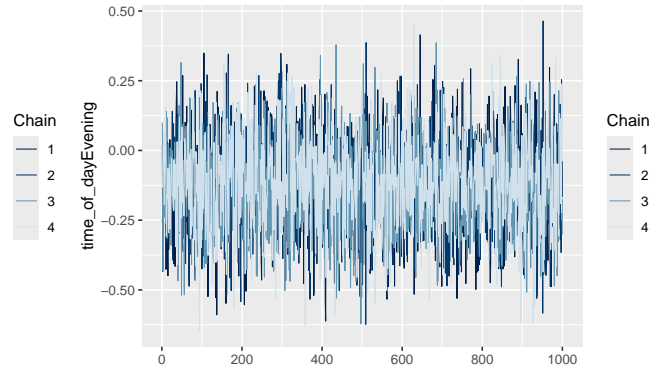
(a) Trace Plot for Intercept



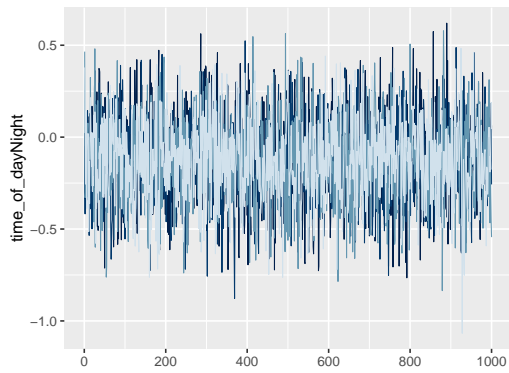
(b) Trace Plot for log Bike Cost



(c) Trace Plot for Morning



(d) Trace Plot for Evening



(e) Trace Plot for Night

Figure 9: Trace plots for time_of_day, intercept, and log_bike_cost

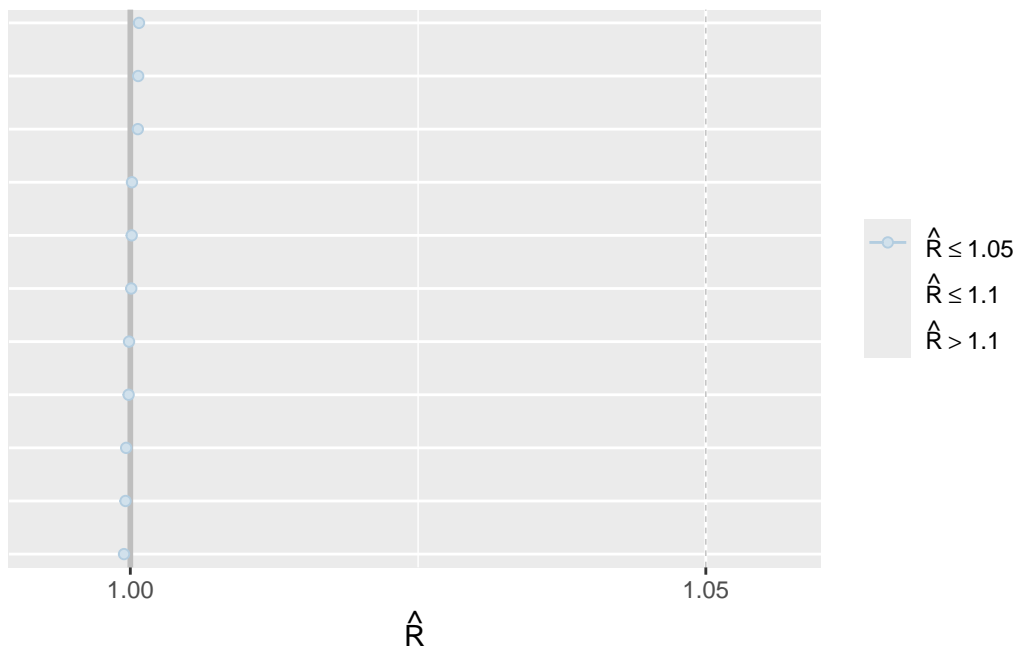


Figure 10: Rhat plot

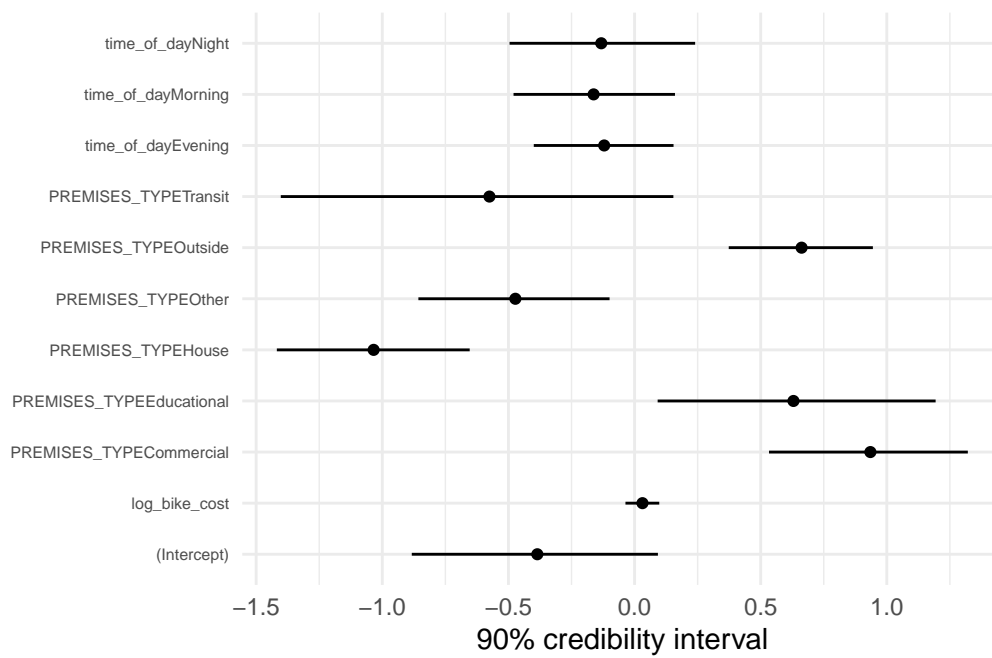


Figure 11: Credible intervals for predictors of positive poverty status

References

- Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. “Joint Longitudinal and Time-to-Event Models via Stan.” https://github.com/stan-dev/stancon_talks/.
- Grolemund, Garrett, and Hadley Wickham. 2011. “Dates and Times Made Easy with lubridate.” *Journal of Statistical Software* 40 (3): 1–25. <https://www.jstatsoft.org/v40/i03/>.
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal* 10 (1): 439–46. <https://doi.org/10.32614/RJ-2018-009>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoş Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. <https://CRAN.R-project.org/package=arrow>.
- Toronto Police Service. 2024. “Bicycle Thefts Open Data.” Toronto, Canada. <https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts-open-data/about>.
- University of Toronto. n.d. “Introduction to GIS Using r.” <https://mdl.library.utoronto.ca/technology/tutorials/introduction-gis-using-r>.
- Wickham, Hadley. 2011. “Testthat: Get Started with Testing.” *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.
- . 2023. *Forcats: Tools for Working with Categorical Variables (Factors)*. <https://CRAN.R-project.org/package=forcats>.
- 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>.
- 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.