DATASCI 451 Final Report

Group 21

Rachel Rubanguka Hoops Xieyao Yin Jeremy Liu Casey Lee

November 2024

1 Introduction

Traffic accidents in the United States are a critical public safety concern with devastating consequences. These incidents not only disrupt lives but also place considerable strain on public resources. Understanding and addressing this issue is crucial for improving road safety and alleviating its impact on individuals and communities (Moosav, 2019).

This study seeks to answer the central research question: What are the key factors contributing to traffic accidents in the United States, and how can we estimate accident probabilities? To explore this question, we focus on environmental and temporal variables that may influence accident occurrence, such as weather conditions, time of day, and roadway characteristics. Utilizing a Bayesian Ordinal Regression model, we aim to estimate accident probabilities under diverse scenarios. This approach uses statistical tools to clearly understand and predict the factors affecting road safety.

2 Mathematical Notation and Model Formulation

Let $y_i \in \{1, 2, 3, 4\}$ represent the severity level of the i-th traffic accident, where: $y_i = 1$ represents a minor accident and $y_i = 4$ represents a fatal accident. The predictor variables

$X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ represent environmental and temporal features of the accident. The probability of each severity level was modeled using Bayesian Ordinal Logistic Regression. The logit function of the cumulative probability is written as

$$\log\left(\frac{P(Y \leq c)}{P(Y > c)}\right) = \alpha_c - X\beta$$

where α_c is the categorical threshold for category c. The weakly informative priors for model parameters are specified as

$$\beta \sim \mathcal{N}(0, 1)$$

$$\alpha_c \sim \text{Cauchy}(0, 2)$$

The likelihood function is based on the ordinal logistic model and it is used, along with the prior distributions of α and β, to estimate the posterior distribution. The posterior is written as

$$P(\beta, \alpha \mid y, X) \propto P(y \mid X, \beta, \alpha)P(\beta)P(\alpha)$$

and the parameters are estimated with the Metropolis-Hastings MCMC algorithm.

3 Data

3.1 Data Download and Features

The dataset was sourced from Kaggle US Accidents (2016 - 2023), comprising 7.7 million traffic accident records across the United States. Each row represents an individual accident, with columns detailing attributes such as location, time, severity, and environmental conditions. The data, available in CSV format, was imported into R for preprocessing and analysis.

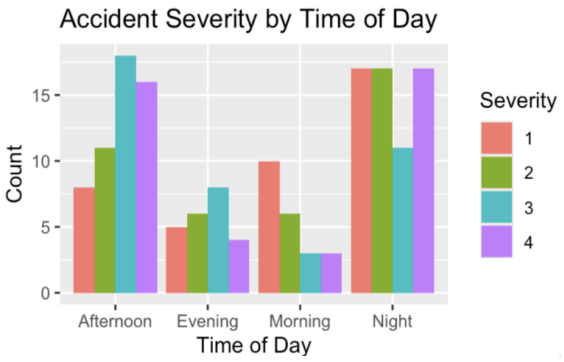
This study focuses on identifying factors contributing to traffic accidents, with key predictors including Weather Condition, Temperature(F), Visibility(mi), and Distance(mi). The target variable,

Severity, quantifies the severity of each accident.

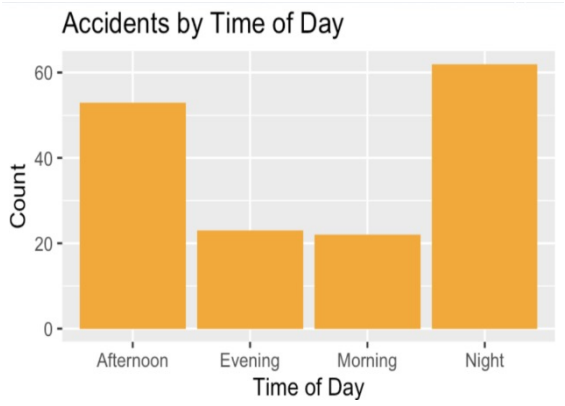
3.2 Data Preview

Temporal and environmental analyses revealed key patterns in accident severity and frequency. Peak accident occurrences were observed at night, likely due to reduced visibility and driver fatigue. Accidents during the night were more severe, with higher proportions of levels 3 and 4, while the morning was dominated by level 1 (minor accidents).

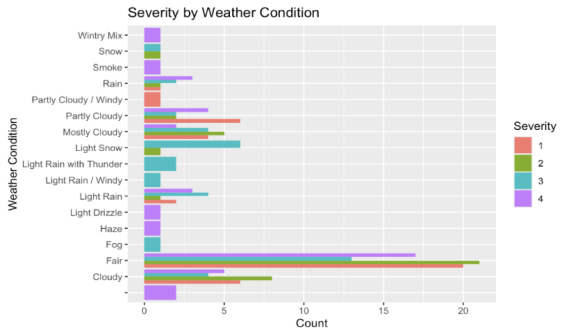
Extreme weather conditions, while less predictive than road-related features, were linked to higher severity accidents. However, fair weather conditions accounted for the majority of accidents across all severity levels, with severity 2 being the most frequent under such conditions.



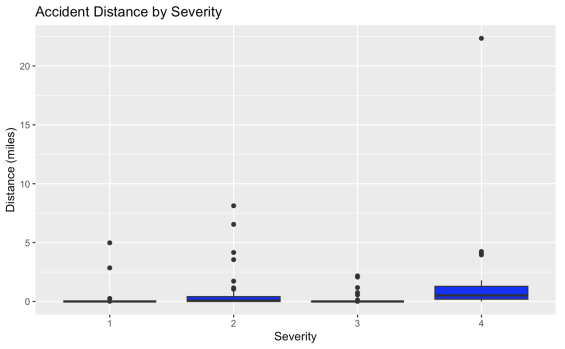
(a) Accidents by Time of the Day



(b) Accidents by Time of the Day



(c) Severity by Weather Conditions



(d) Accident Distance by Severity

Figure 1: Analysis of Accident Severity by Time of Day, Weather, and Distance

3.3 Data Cleaning and Pre-Processing

The data was restricted to relevant predictors, ignoring redundant time, location, and accident description columns. Columns missing over 50% of their values and rows missing key predictor values were omitted. This helped reduce data sparsity while retaining essential observations.

The Start Time and End Time columns were converted to datetime objects and the accident duration was calculated in minutes. Categorical columns, such as Weather Condition, were consolidated for analytical consistency. A Time of Day column was created by grouping accidents into time periods (Morning, Afternoon, Evening, Night) according to the accident's starting hour. The cleaned dataset was saved as a CSV file for reproducibility and sampling.

Binary predictors (such as those that addressed whether there was an amenity, bump, or traffic signal in the traffic accident) were encoded as integer values. Continuous predictors were converted to numerics and were standardized using z-score normalization. Categorical predictors were encoded as integer factors.

Due to mechanical limitations, the study used a subset consisting of 400 sampled entries per Severity level to equally represent minority categories.

4 Analysis

4.1 Methodology

This study used a Bayesian ordinal regression model to predict traffic accident severity, leveraging predictors related to the weather, road conditions, and time. The model was chosen due to the ordinal nature of the severity response variable.

In order to analyze the true contribution of individual predictors, predictors that exhibited high collinearity were omitted, namely several of the dataset's multiple weather predictors (e.g. Humidity and Visibility). Different interaction terms were tested to examine more nuanced relationships between predictors and their combined influence in predicting accident severity. From there, ridge

regression was run against the ordinally factored variable Severity with the goal of identifying and filtering out the most predictive features.

A Bayesian logistic regression model was also implemented as a comparative approach, treating severity levels as nominal categories. However, it underperformed due to the loss of ordinal structure, supporting the appropriateness of the Bayesian ordinal regression framework.

4.2 Predictors Used

Cross-validated Lasso regression identified the most predictive variables, including *Temperature.F.*, *Pressure.in.*, *Weather_Condition*, *Wind_Speed.mph.*, and *Crossing*, which were consistently significant across severity classes. By applying regularization to the model, these predictors were selected based on their non-zero coefficients at the optimal penalty parameter ($\lambda = 0.0045$).

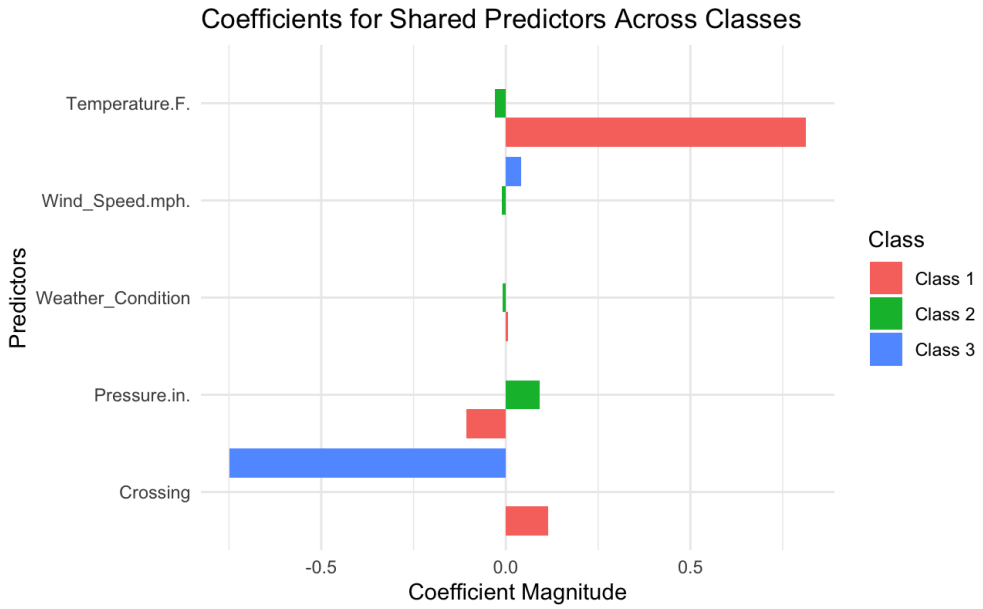


Figure 2: Coefficients of Shared Predictors Across Severity Classes from Lasso Regression

Ridge regression provided an overview of a feature's predictive relevance, but further tests with different combinations of predictors and their interactions were necessary to properly assess their individual contributions and to prevent model overfitting.

Through the process of including all pairwise interaction terms and filtering by correlation and the

magnitude of their regularized coefficients, it was found that including interactions between the binary features improved the classification rate.

In the final model, the data was loaded, scaled, and cleaned by selecting the most predictive variables and converting them into numerics. A matrix was created for predictors, and a vector was created for the response variable, Severity, for model fitting. The chosen predictors are as follows:

1. **Binary:** Crossing, Amenity, Traffic Signal, Junction.
2. **Continuous:** Temperature, Distance, Wind_Speed.mph, Precipitation.in.
3. **Interactions:**
    - Crossing * Traffic Signal,
    - Traffic Signal * Amenity,
    - Crossing * Amenity,
    - Amenity * Traffic Signal * Crossing.

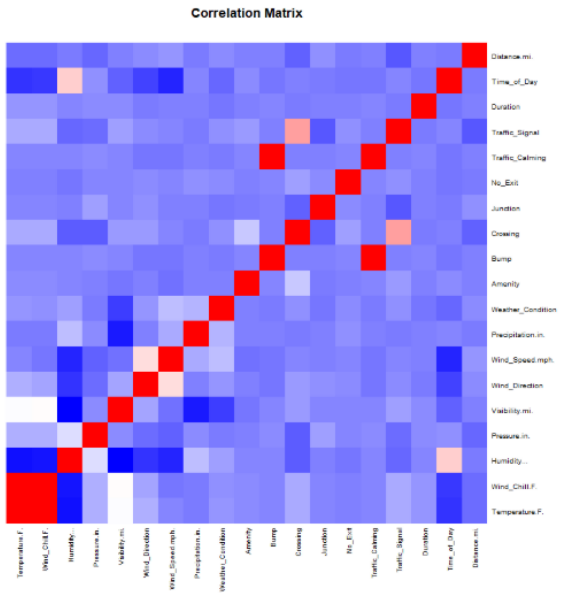
5 Findings and Simulation

5.1 Predictor Importance

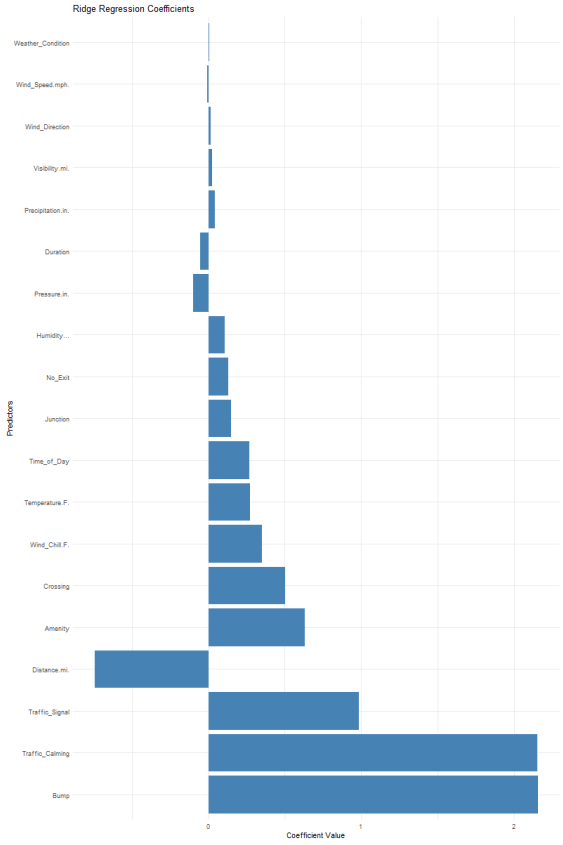
Road-related predictors, such as Traffic Signal and Junction, were found to be more influential than weather-related variables and resulted in higher classification rates and lower mean absolute errors. Ridge regression results indicated the lower predictive power of weather variables relative to the binary variables (Figure 3b).

5.2 Model Performance Results

The Bayesian ordinal regression model demonstrated reasonable predictive accuracy, with predicted severity counts aligning closely with observed counts. The posterior predictive plot visually

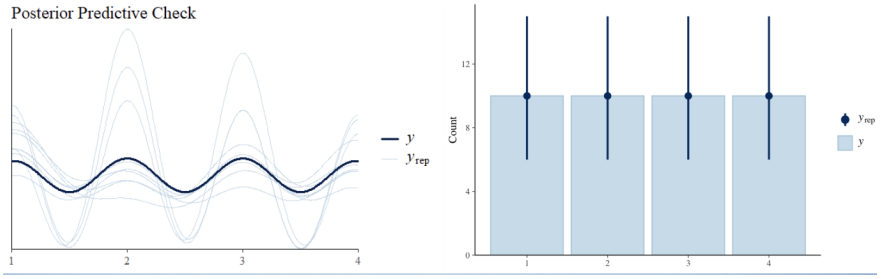


(a) Correlation heatmap of the filtered predictors. Blue represents a negative correlation and red positive.



(b) Initial ridge regression run against Severity to gauge predictor performance.

Figure 3: Comparative visualizations: correlation heatmap and ridge regression coefficients.



(a) Posterior Predictive and Predicted Counts

confirms this alignment, showing predicted counts centered around observed counts with some uncertainty intervals. The predicted counts (y rep) are roughly centered around the observed counts (y) for each severity level. From the graphs, we can see that the uncertainty intervals are relatively large.

5.3 Bayes Ordinal Model: Data Fitting Results

The metrics R_{hat} and Effective Sample Size (ESS) were utilized to confirm model convergence:

- The parameters converged well, with a minimum ESS of 2800 across all chains.
- R_{hat} values were very close to 1.

Evaluation metrics included misclassification rates and predictions' Mean Absolute Error (MAE):

- Misclassification rate: 0.635.
- Mean Absolute Error (MAE): 0.91.

Despite the misclassification rate not significantly outperforming the naive prediction model misclassification rate of 0.75, predictions typically remained within one category of the actual severities. This result indicates reasonable performance in predicting accident severity.
Density plots showed reasonable credible intervals for the majority of the original predictors' coefficients, though the coefficient for Amenity is evidently less reliable. The wider credible intervals for Amenity and the interaction terms can both be attributed to a low sample size; the interaction terms suffer especially so as they only refer to accidents with overlaps between the binary predictors.

The category thresholds α_i had tighter credible intervals, which shows the model's ability to distinguish between Severity categories. However, this contradicts with the high misclassification rate. The relatively small number of samples likely impacted the representation of some predictors, as a predictor's sampled values may not accurately represent the population. Thus, the number of samples used would make it difficult for the model to accurately distinguish between the categories.

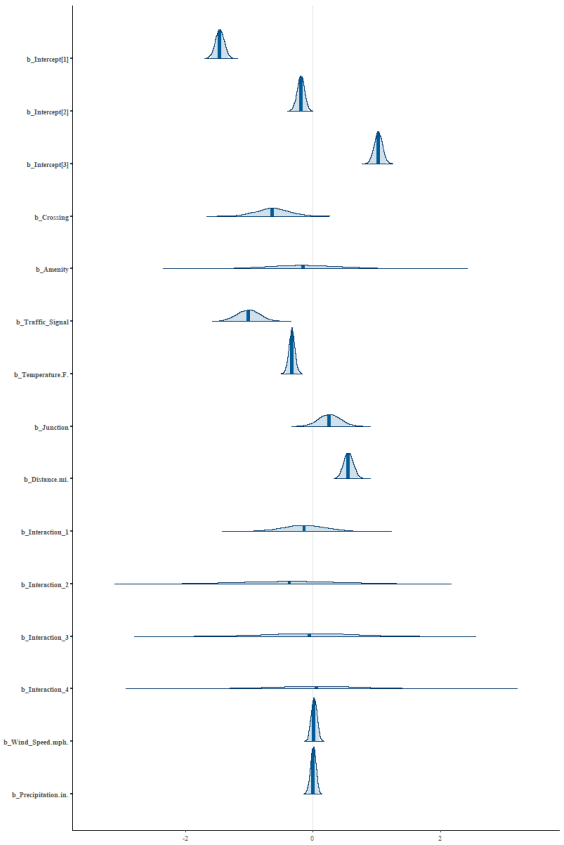
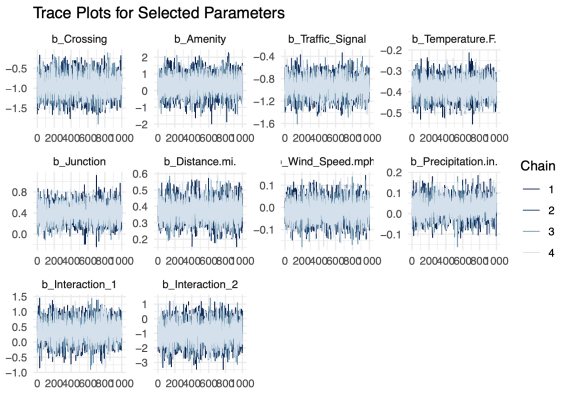


Figure 5: Posterior density plots for the chosen predictors at a 95% credible interval.

5.4 MCMC Convergence



(a) MCMC Convergence

Trace plots revealed stable posterior estimates for most predictors, indicating successful MCMC convergence. Again, some parameters exhibit more fluctuations that may indicate higher levels of uncertainty for the posterior estimates.

6 Conclusion

This study demonstrates the effectiveness of Bayesian ordinal regression in predicting traffic accident severity, providing a reliable framework for analyzing key contributing factors. Road-related predictors along with interactions between the binary road-related predictors emerged as the most influential variables in predicting accident severity, outweighing weather-related factors in predictive importance. While the model achieved reasonable performance, challenges such as overlapping feature distributions and sampling inaccuracies persist. The lack of samples also impacted the representation of minority subcategories (e.g. Southwest in the WindDirection predictor), making it difficult for the model to properly draw conclusions. These issues were difficult to mediate due to mechanical limitations. Future work could address these issues by utilizing a larger sample size in order to better emphasize the relationships between the predictors and the response as there are inherent issues in randomly drawing such a small sample relative to the original dataset. In the end, the study's findings underscore the importance of considering infrastructural factors in accident prevention strategies, as weather-related predictors were found to be less of a factor in predicting crashes compared to road-related predictors.

7 Bibliography

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.