Car Crash Reporting Prediction Based on Poisson and Negative Binomial Bayesian Model

Sifan Tao, Jingyun Jia, Xinyan Wang

April 27, 2024

1 Introduction

Nowadays, road safety is attracting more and more attention. Inspired by that, we are interested in crashing reporting prediction.

Exploring crash reporting prediction can be valuable from many aspects - firstly, by predicting crashes and understanding their causes, measures can be taken to prevent them, leading to safer environments for individuals; secondly, for optimized maintenance and operations, predictive analytics can guide maintenance schedules and operational strategies, reducing downtime and ensuring that road systems remain in optimal working condition.

The rest of this paper is organized as follows: Section 2 introduces our raw data structure, data pre-process, cleaned-data structure, variables, and explanations; Section 3 explains our main models; Section 4 shows our empirical results; Section 5 draws conclusions and discusses the potential improvement of models.

2 Data

Our dataset reports details of all traffic collisions occurring on county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police. The data can be found at https://catalog.data.gov/dataset/crash-reporting-drivers-data.

The raw dataset contains a total of 172105 data points and 43 variables, and each data point is a crash record. Since many variables are duplicated or unrelated to prediction, we select 4 variables - route type, surface condition, light condition, and whether the driver's license is from Maryland or not. After combining some categories within each variable, we get bar plots Figure 1 and Figure 2 showing their distributions.

From these two figures, we can see that most crashes happened on highway and Maryland state routes, on dry roads in the daylight, with the drivers having Maryland driving licenses.

We set the response to be the number of crashes under each combination of variables. After summarizing, there are 146 different combinations and therefore, we have 146 responses. However, since we have 5, 5, 4, 2 levels for 4 predictors, correspondingly, the number of full combinations is 200. We complete the dataset by setting the responses of those missing combinations to be 0. Eventually, the dataset we use contains 200 observations, each raw represents a unique combination with the number of crash accidents. In addition, all predictors are categorical variables, we used **fastDummies** by Kaplan (2023) to create dummy variables for each predictor.

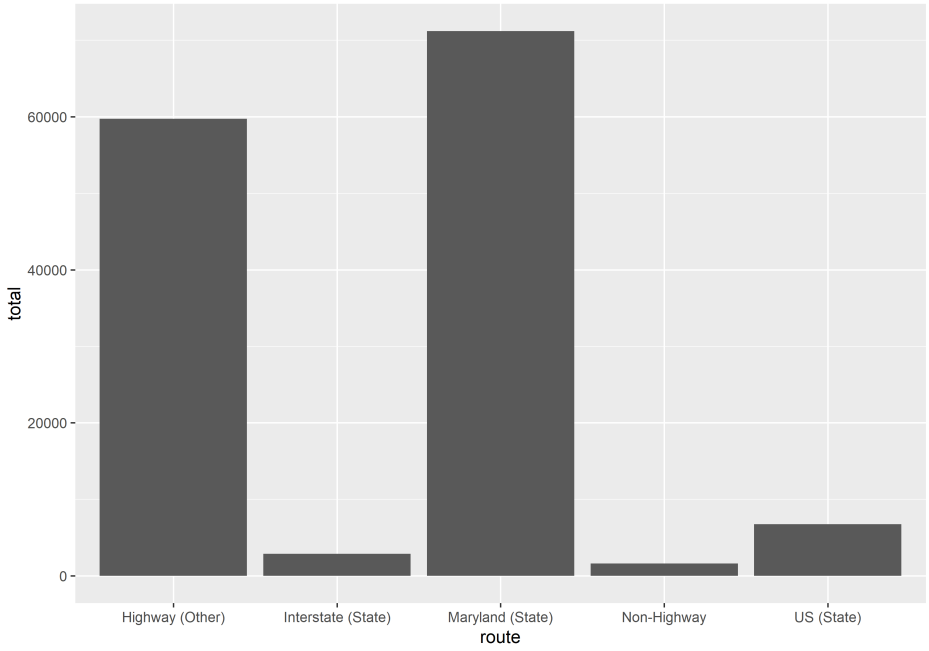

Figure 1: Distribution of Route Types

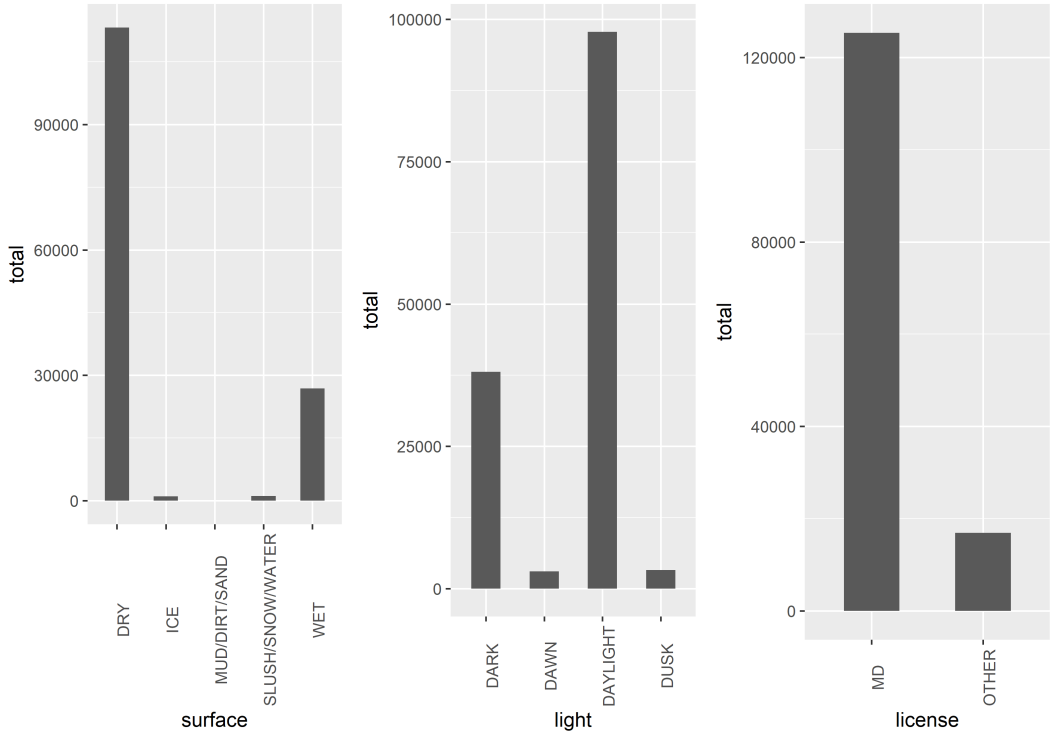

Figure 2: Distribution of Surfaces, Light and Licenses

3 Models

In this section, we will first introduce an initial model and then develop two refined models, the Poisson Model with Offset and the Negative Binomial Model with Offset based on the initial model.

3.1 Initial Model

After summarizing, the responses Y are the number of car crash accidents under certain conditions. For example, Y_i could be the number of car crash accidents that happened on a dry, US(State) route in dark light, and the driver holds a Maryland driving license. We assume that $Y_i's$ are the number of independent events occurring in a fixed time period. Therefore, it is natural to fit a Poisson model. Specifically, the initial model(Naive Poisson Model) is:

$$Y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$\beta_{j[k]}|\beta_j, \sigma \sim N(\beta_j, \sigma^2)$$

$$\beta_j|\tau \sim N(0, \tau)$$

$$\alpha|\mu_\alpha, \sigma \sim N(\mu_\alpha, \sigma^2)$$

$$\mu_\alpha \sim N(0, \tau)$$

$$\sigma^2 \sim IG(3, 3)$$

$$\tau^2 \sim IG(3, 3),$$

where $\lambda_i's$ are the rate of number of car crash accidents, α is the overall baseline following a zero-mean normal distribution, $\beta_j's$ are coefficient vectors for different predictors, Surface Condition, Light Condition, License Type, and Route Type, $j = 1, 2, 3, 4$ correspondingly. We assume all random variables mentioned above are independent. In addition, because different levels $\beta_j[k]$ of the same predictor should show some weak similarity, we assign a common mean β_j for different levels. Then, we discuss how to refine the naive Poisson model.

3.2 Refined Models

With all 4 explanatory variables in hand, the naive model treats them all as predictors. However, the variable route type may not have a direct influence on the response and is highly correlated to other variables. For example, most accidents happen on Maryland State routes and other highways due to the large volume of those types of routes and the relatively high speed on those routes. According to Wikipedia (2022), there are 16 interstate highways, 14 U.S. highways, 998 state highways labeled from 2 to 999, and other routes.

Because we do not have access to the number of other routes, it is infeasible to add an offset use the above information. Therefore, we settle for second-best and consider the number of accidents happening on different routes as an offset. Our first refined model, namely the Poisson Model with Offset, is proposed with the same notation as in the Naive Poisson Model:

$$
\begin{aligned}
Y_i &\sim Poisson(\lambda_i) \\
log(\lambda_{i,route}) &= \alpha + log(N_{route}) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
\beta_{j_{[k]}} | \beta_j, \sigma &\sim N(\beta_j, \sigma^2) \\
\beta_j | \tau &\sim N(0, \tau) \\
\alpha | \mu_\alpha, \sigma &\sim N(\mu_\alpha, \sigma^2) \\
\mu_\alpha &\sim N(0, \tau) \\
\sigma^2 &\sim IG(3, 3) \\
\tau^2 &\sim IG(3, 3),
\end{aligned}
$$

where $log(N_{route})$ is the total number of accidents on a certain kind of route in log scale and the information is summarised in Table 1. As the information of route types has already been treated as offset, we remove it from the predictors and there are 3 explanatory variables in the model.

Table 1: Total Number of Accidents on Different Routes(in log scale)

Route	Highway (Other)	Interstate (State)	Maryland (State)	Non-Highway	US (State)
log(N)	10.998	7.973	11.173	7.386	8.821

In addition, the Negative Binomial model is also a popular approach to fit count data. In general, the Negative Binomial model is more flexible than the Poisson model because we need to assume that variance and mean are equal in the Poisson model, which is not required in the Negative Binomial model. Therefore, another refined model called Negative Binomial Model with Offset is as follows:

$$Y_i \sim NegativeBinomial(\lambda_i, \phi)$$

$$\phi = \frac{1}{\eta^2}$$

$$\eta \sim N(0, 1)$$

$$log(\lambda_{i,route}) = \alpha + log(N_{route}) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\beta_{j_{[k]}} | \beta_j, \sigma \sim N(\beta_j, \sigma^2)$$

$$\beta_j | \tau \sim N(0, \tau)$$

$$\alpha | \mu_\alpha, \sigma \sim N(\mu_\alpha, \sigma^2)$$

$$\mu_\alpha \sim N(0, \tau)$$

$$\sigma^2 \sim IG(3, 3)$$

$$\tau^2 \sim IG(3, 3),$$

where ϕ is the dispersion parameter with a relatively flat prior. The detailed notations in two refined models are summarised in Table 2.

Table 2: Notations of the Parameters

Factor	Level	Parameter
Surface	dry	$\beta_{1[1]}$
	ice	$\beta_{1[2]}$
	mud/dirt/sand	$\beta_{1[3]}$
	slush/snow/water	$\beta_{1[4]}$
	wet	$\beta_{1[5]}$
Light	dark	$\beta_{2[1]}$
	dawn	$\beta_{2[2]}$
	daylight	$\beta_{2[3]}$
	dusk	$\beta_{2[4]}$
License	Maryland	$\beta_{3[1]}$
	non-Maryland	$\beta_{3[2]}$

4 Empirical results

In this section, we will compare two refined models - the Poisson Model with Offset, and the Negative Binomial Model with Offset with different criteria.

4.1 Experiment Setting

With the specified models in Section 3, we used **RStan** to fit the Poisson Model with Offset and the Negative Binomial Model with Offset. For both models, we standardized the α and $\beta_{j_{[k]}}$ to prevent sampling issues during Hamiltonian Monte Carlo Neal (2003). After the reparametrization, we have

$$\beta_{j_{[k]}}|\beta_j, \sigma = \beta_j + \sigma\beta_{j[k],aux}, \beta_{j[k],aux} \sim N(0,1),$$

$$\alpha|\mu_\alpha, \sigma = \mu_\alpha + \sigma\alpha_{aux}, \alpha_{aux} \sim N(0,1)$$

The standardized parameters are denoted by a subscript 'aux' and a detailed explanation is in Table 3.

Table 3: Notations of the Standardized Parameters

Factor	Level	Parameter(after Standardization)
Surface	dry	$\beta_{1[1],aux}$
	ice	$\beta_{1[2],aux}$
	mud/dirt/sand	$\beta_{1[3],aux}$
	slush/snow/water	$\beta_{1[4],aux}$
	wet	$\beta_{1[5],aux}$
Light	dark	$\beta_{2[1],aux}$
	dawn	$\beta_{2[2],aux}$
	daylight	$\beta_{2[3],aux}$
	dusk	$\beta_{2[4],aux}$
License	Maryland	$\beta_{3[1],aux}$
	non-Maryland	$\beta_{3[2],aux}$

During the sampling process, we set the initial value to be 0 for each chain with 20000 iterations.

4.2 Model Evaluation

After fitting the Poisson Model with Offset and Negative Binomial Model with Offset, we first plotted the convergence graph for $\beta_{1_{[k]}}$, the coefficient for different surfaces for both models. Both models showed a good convergence on the $\beta_{1_{[k]}}$ and both models yield a similar value for each $\beta_{1_{[k]}}$ as shown in Figure 3. We also compare the posterior predictive of the rate $\lambda|y$ against the observed rate λ in Figure 4 and Figure 5.

For the Poisson Model with Offset, the 95% credible intervals for the posterior predictive distribution include 128 out of 200 of the rate λ, while the Negative Binomial Model with Offset includes 153 out of 200 of the rate λ. We noticed that the Negative Binomial Model includes more λ due to its large variance, which leads to a larger credible interval compared to the Poisson Model. From the perspective of the posterior predictive of the rate λ, we would prefer the Negative Binomial Model with Offset over the Poisson Model with Offset. We then computed the relative root mean square error of the posterior predictive counts $y|Y_i$ for both models, shown in Table 5. To be noted, because all the predictors are categorical variables, the models will have limitations on fitting the data and the acceptance rate will be less than the expectation correspondingly.

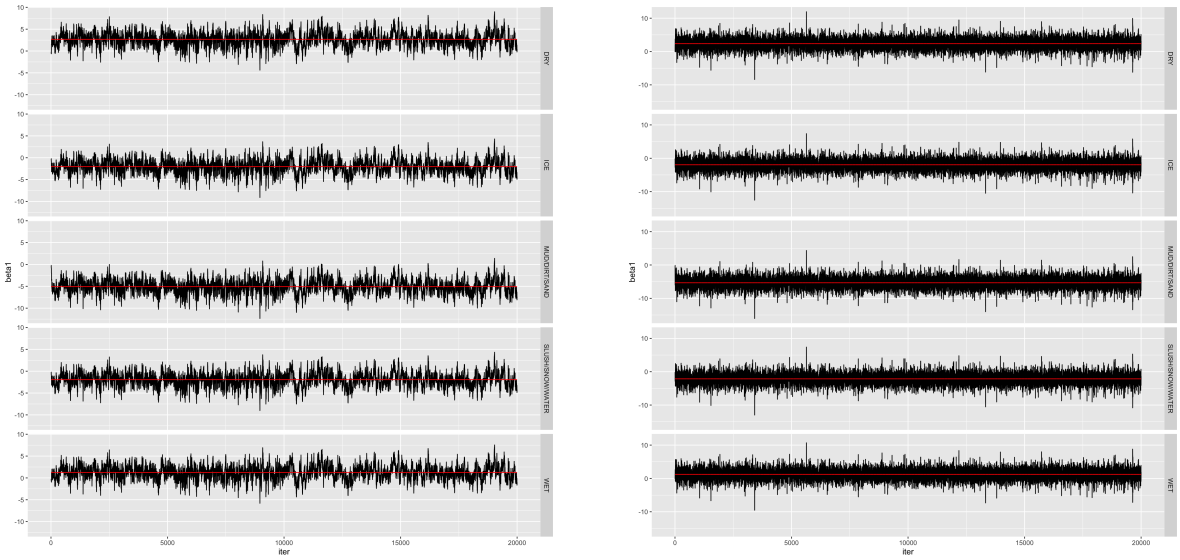

Figure 3: Convergence Graph of β_1 for Refined Models(Left is Poisson and Right is Binomial)

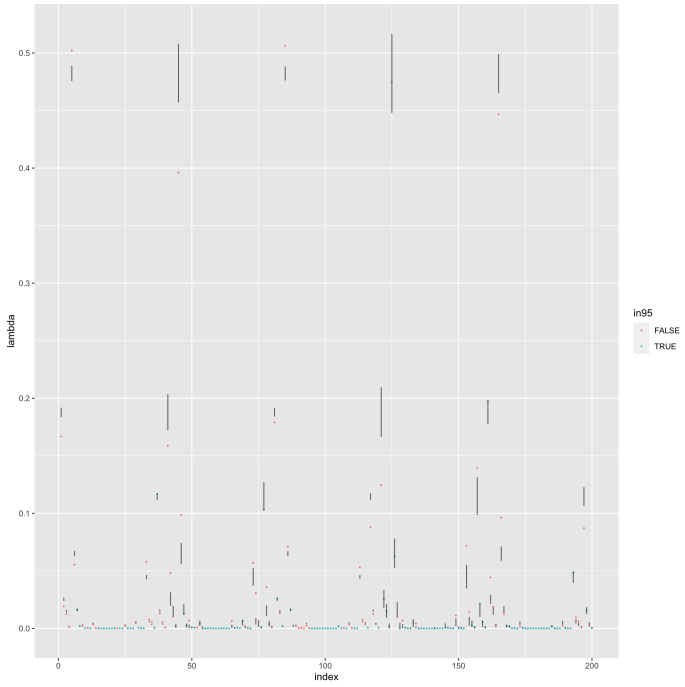

Figure 4: 95% Credible Interval for Posterior $\lambda|y$ for Poisson Model with Offset

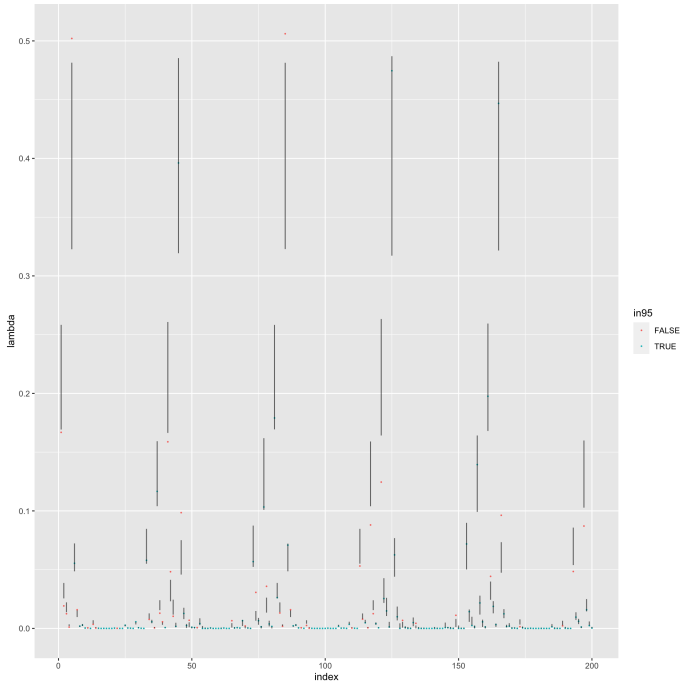

Figure 5: 95% Credible Interval for Posterior $\lambda|y$ for Negative Binomial Model with Offset

Table 4: Model Evaluation for Poisson and Negative Binomial Models

Criterium	Poisson with Offset	Negative Binomial with Offset
RRMSE	0.005	0.018
y_{max}	0	0.008
y_{min}	0	0
λ_{max}	0.196	0.016
λ_{min}	0	0

The Poisson Model with Offset has a much smaller RRMSE than the Negative Binomial Model with Offset. The posterior prediction from the Poisson Model with Offset is much closer to the actual count data observed, but they have a smaller variance compared to the Negative Binomial Model with Offset. This relationship is very similar to the graph shown above for the posterior predictive of rate λ. The RRMSE would prefer the Poisson Model with Offset instead of the Negative Binomial Model with Offset, which yields a different conclusion from the posterior predictive. We also checked the distribution of the posterior predictive for the count $y|Y_i$ and compared it with the observed counts Y_i.

We can also evaluate the model using Bayesian p-valueMeng (1994). Bayesian p-value allows us to use posterior predictive information of the data. We compare the minimal and the maximal statistics of the observed y(and rate) and the predicted y(and rate) and compute the tail probability that the predicted quantities are larger than the observed ones. The results are in Table 5. Because our observed data includes multiple 0', the minimal of the predicted y(and rate) is 0 for all simulations. And we can also see that the predicted values in the Poisson Model tend to be much larger for the max(rate).

Moreover, we can check the distribution of the predicted y and the observed y. From the histogram in Figure 6, we can see that the Poisson model tends to shrink the values that are small.

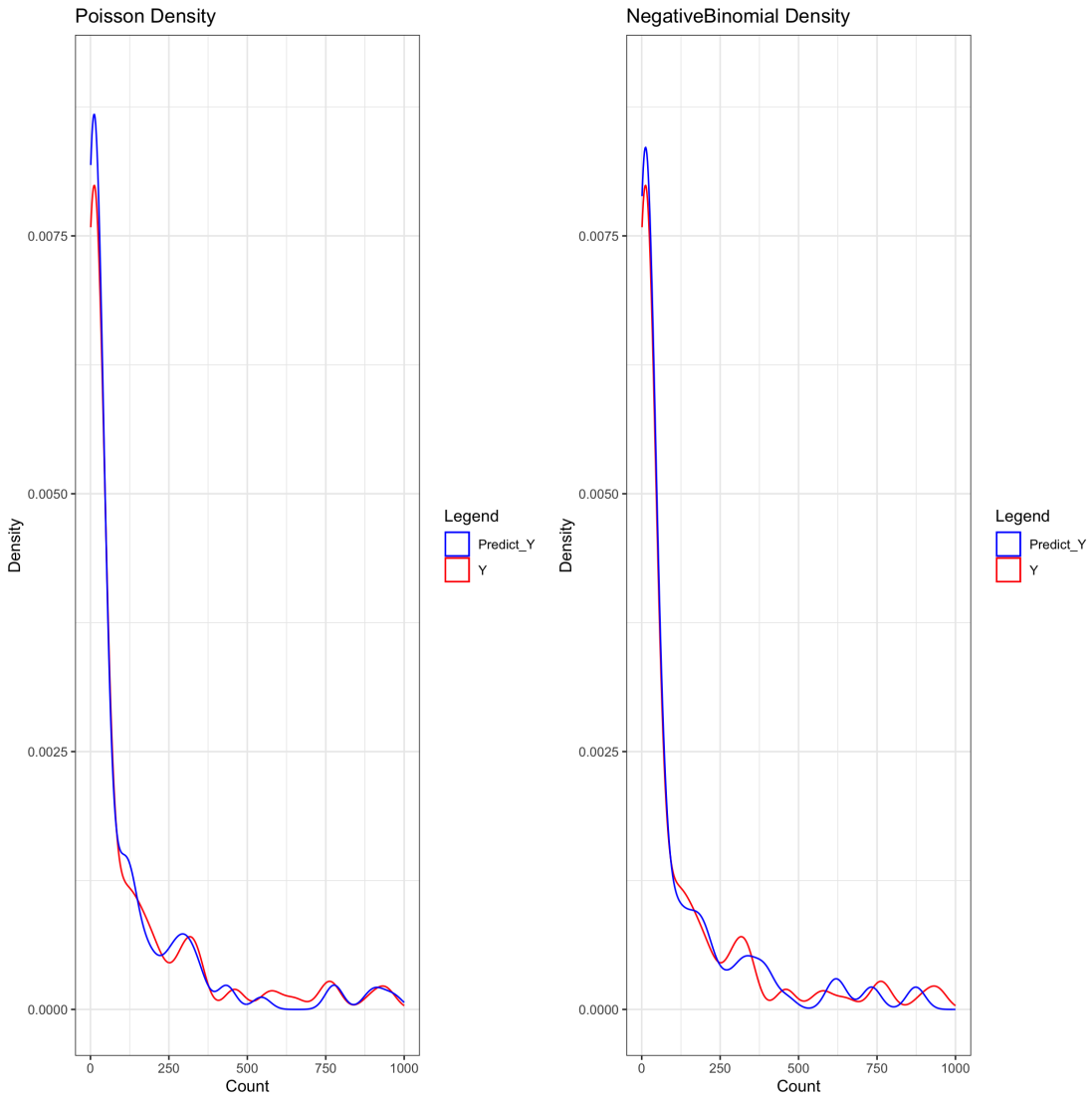

Figure 6: Distribution of the Predicted y and the Observed y in Poisson and Negative Binomial Models

4.3 Model Interpretation

The posterior estimation of the parameters for both the Poisson Model with Offset and the Negative Binomial Model with Offset are shown in the table below.

Table 5: Posterior estimation of the Standardized Parameters for Poisson and Negative Binomial Models with offsets

Parameter	Poisson with Offset	Negative Binomial with Offset
$\beta_{1[1],aux}$	1.507	1.462
$\beta_{1[2],aux}$	-0.683	-0.570
$\beta_{1[3],aux}$	-2.098	-2.164
$\beta_{1[4],aux}$	-0.642	-0.656
$\beta_{1[5],aux}$	0.839	0.933
$\beta_{2[1],aux}$	0.242	0.303
$\beta_{2[2],aux}$	-0.936	-0.867
$\beta_{2[3],aux}$	0.681	0.602
$\beta_{2[4],aux}$	-0.897	-1.035
$\beta_{3[1],aux}$	-0.072	-0.061
$\beta_{3[2],aux}$	-1.004	-0.955

The coefficients of the two refined models are similar. We explain our results that the variables whose coefficients have large values of posterior estimation will have more crashes. This corresponds to what we have shown in the EDA section - the number of crashes under each level of variables. For example, in the EDA part, most crashes happen when the surfaces are dry; our posterior estimation of the coefficient of the dry surface is also the largest among all other levels of surface condition. That means the structure of our models is reasonable and the models have great capacity in terms of predicting the number of crashes if the external conditions are given.

5 Discussion

In summary, the Negative Binomial model with offsets is better in the sense of the higher acceptance rate, whereas the Poisson model with offsets is better in the sense of the smaller RRMSE; both of our refined models can provide an effective prediction of the number of crashes in Montgomery County under different external conditions.

In addition, we also tried the zero-inflated Poisson model and zero-inflated Negative Binomial model, as the data has a extremely large amount of zero observations. However, the acceptance rates for both zero-inflated models are smaller than the presented two refined models. For future work, we can adjust the prior of zero-inflated models and consider involving the truncated distribution.

References

Kaplan, J. (2023). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. R package version 1.7.3.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705 – 767.

Wikipedia (2022). Maryland highway system.