Bayesian Analysis of New York Car Crashes

Jake Ryan, Aditya Fuldeore, Jefferson Mathews, David Zhu

Introduction

In this project we will use a set of descriptive analysis and prior analysis to examine the Bayesian analysis of the New York Car Crashes. The data is reported by police reports, rounded to the nearest hour, and if the time is unsure, they put it at midnight. This might cause some discrepancies between the actual dataset. We will also be rounding our time variable to the nearest hour because in police reports it is very common to round the time to the nearest hour, so our data has some values that are rounded off and some that are exact. To make our data more uniform we decided to round all of our times to the nearest hour. We will also be applying a variety of modeling and graphs to examine our data, and the prior analysis involves a generalized linear model analysis, as well as time series analysis.

Upon initial looks, the data had large crash numbers that range from hour 9 to 19. The methods used include the directed acyclic graph, as well as the JAGS and model graphs to study the crash data. A non informative prior is applied into the data analysis and also examines how different parameters play a different role into the effect of the data and the model building process, and compare the different results with an informed prior. We are also going to split the data up into different sections of time that represent early hours, mourning rush, afternoon rush, and nighttime. Based on our preliminary analysis there was a distinct difference in the number of accidents given these different times of day, as well as a distinct difference in how the rate of accidents were changing over these time periods.

Methods

The Bayesian hierarchical model was constructed with the likelihood calculated as the average number of accidents in an hour according to a poisson distribution with parameter lambda. This was chosen due to the time nature of the data that was being answered. The data is a rate per hour dataset in context, so the analysis of it corresponded to this trait. The priors were noninformative betas one and two that were distributed according to a normal distribution with a low precision. These priors were the most effective for ensuring convergence after a minimal amount of burn-in. Using an informed prior with different parameters most likely would have slowed the model building process and not given an effective model for the purposes of this

report. After constructing the hierarchical model, the JAGS code as found on pg 8-9 in the pdf in the appendix was used for calculations. This gave the corresponding results that are discussed later in the report.

To ensure convergence in the model, the betas on the three chains were set to values that would be very separate from each other initially. These values were -1, 0, and 1 respectively. As shown by the gelman plots on page 10 of the pdf in the appendix, these values did ensure the three chains converged in the model. Which means the separate chains of the model all end up at the same target distribution. This convergence indicates a stronger model and gives greater credibility to the use of the model in explaining the relationship between time of the accident and the rate. The x and y variables in the code were used in the calculations of lambda in the model. X was the number of accidents and Y represents a categorical variable that marked the hour of the accident 0-23 corresponding to the hour in the day that it occurred. The mean of the X's were calculated according to the corresponding Y's to give total accidents in our data in each hour. Then these values were distributed according to a poisson distribution with parameter lambda that was calculated for an hour of the day. The mean values of the lambda, beta 1, and beta 2 were then the results.

The size of the data set allowed for the use of non-informative priors. This made the initial model building more efficient. However, informed priors could be acquired by analyzing traffic data to give greater weight to areas that experience accidents at a different rate, such as breaking priors down by borough. As mentioned before, this would be a difficult task to perform and might necessitate its own analysis. There is no obvious weighting that makes sense in terms of an informed prior, so convergence may not occur even if the prior makes sense to use. In this case, the model would be much weaker and hinder the report. Overall, a non-informed prior was the best way to go given the conditions that we were under for this report.

Results

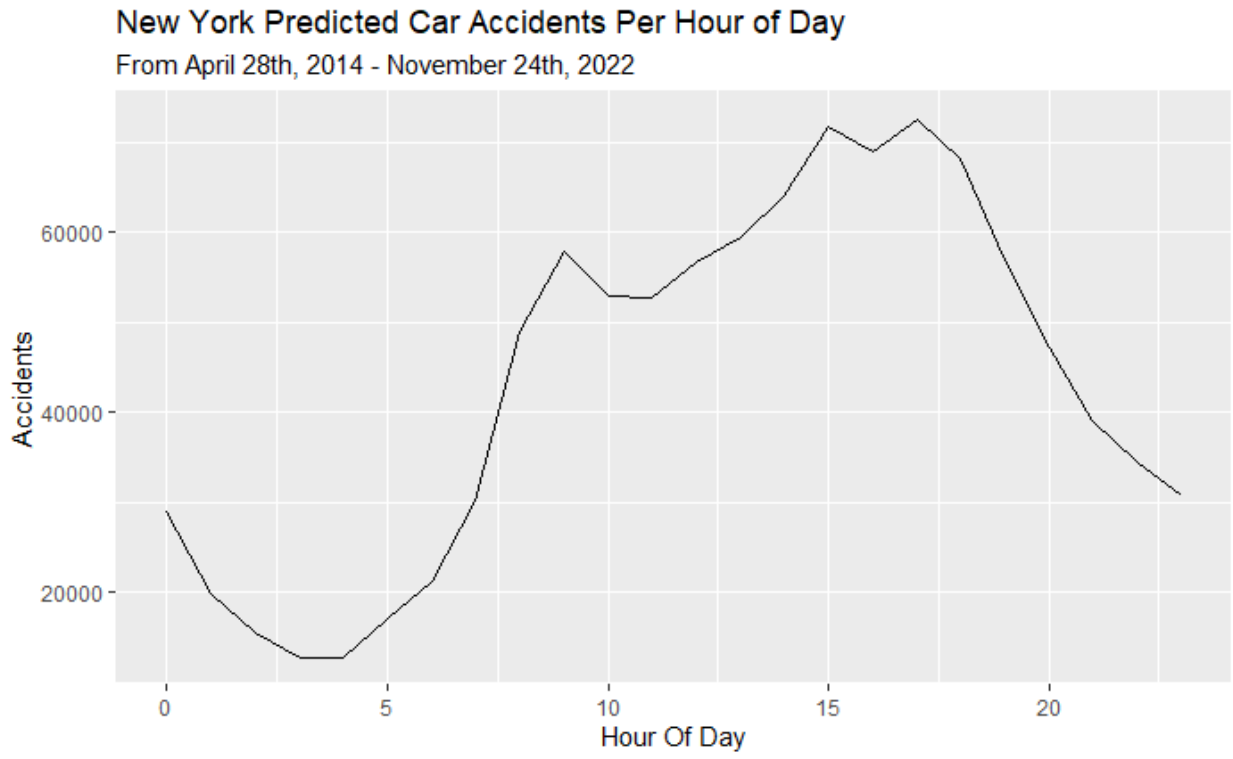
After an analysis of the dataset, we found a result on the rate of car accidents. Our dataset covered accidents from April 28th, 2014 to November 24th, 2022. So, we were able to build a Bayesian hierarchical model that predicted the accidents that occurred within each hour block over that time.

We created four models, running markov chains for different time blocks. These models lead to a few key results, and we need to evaluate to what extent it can be trusted. All four

models converged, with potential scale reduction factors all reaching 1. The models' Deviance Information Criterion (DIC) were also low, no model had a DIC over 45, or a penalty level over 2.1. So, the models had low deviance, meaning they were a better fit than not. The Monte Carlo Errors for each hour were also low, with the highest error of 0.01079 coming at hour 15 (3:00 PM). The full Monte Carlo Errors for each hour can be found in the appendix. The time blocks, DIC, and penalties of each model are shown in the table below. Midnight (12:00 AM) is hour 0, while hour 23 is 11:00 PM.

Hours in Model	**DIC**	**Penalty**
0 - 4	15.26	2.042
5 - 10	27.48	2.064
11 - 15	25.85	1.894
16 - 23	42.56	1.971

Since we have a log transformation of the response in our JAGS code, our posterior values need to be calculated through that transformation. The midnight hour is shown to have a posterior number of 29052.92 accidents. However, this number may be confounded because of how accidents are reported. When the exact hour/time of an accident is unknown, it is reported in our data with an hour of "00", which is the midnight hour. So, the number of predicted accidents at midnight may be confounded because of how accidents are reported in our data. The following plot shows the predicted car accidents from April 28th, 2014 to November 24h, 2022 for each hour of the day, starting at 0 for midnight and ending at 23 for 11:00 PM.



As the plot shows, the predicted accidents in a certain hour reach above 70000 accidents on the y-axis. The actual numerical values of the accidents in an hour are shown in the appendix.

The largest peak in accidents can be seen going from hour 15 to hour 18, or 3:00 PM to 6:00 PM. This is likely because more people are driving in the afternoons and evenings, coming home from work, going shopping, or running other errands and business throughout New York. The 5:00 PM accidents were the highest predicted at 72597.31, which is just 23.18 accidents per day in our 3,132-day dataset. The early morning accidents were low, with 3:00 AM being the lowest block, at 12765.06 accidents, or 4.08 accidents per day for our dataset. The number of accidents picked up at 8:00 AM, when general businesses tend to open up.

A frequentist analysis we had performed in the preliminary stage of the data showed a high AIC, meaning the frequentist method was an ok fit. We used an ARIMA model due to our data being time series, looking at car accidents over the course of each hour of the day. The model gave an AIC value of 471.51, and all but one of the model variables were significant, so it was an ok fit. Compared to prior generalized linear model analysis that we did, the ARIMA time

series model was the best fit, with the lowest AIC. When compared to our Bayesian model, we believe the Bayesian model is a better method due to its use of priors. Both the Bayesian and frequentist methods were solid fits, but the Bayesian model offers better posterior predictability.

Conclusion

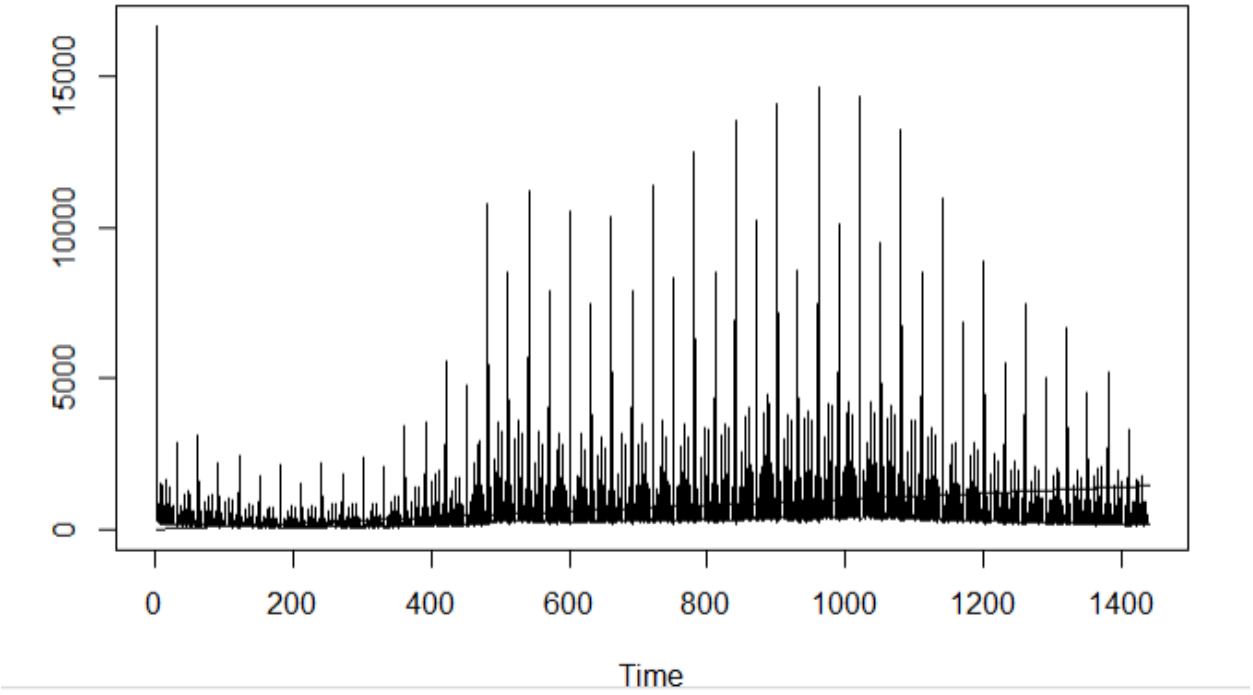
To conclude, we created Bayesian hierarchical models with non-informative priors to predict the number of car accidents in each one-hour block of the day from April 28th, 2014 to November 24th, 2022. The models found that the predicted number of accidents was highest between 3:00 PM to 6:00 PM. Multiple models were needed to best fit the data, and this helped get better posterior predicted accidents over time throughout the day. Ultimately, the highest rate of car crashes occur in the afternoons and evenings, while early mornings have a low rate, likely due to the number of car traffic at those times.

Appendix

- Link to PDF file of the R code and output

https://drive.google.com/file/d/1QboyUP9Za1RYicLV2Zn6SET0vzL2nt7w/view?usp=share_link

- Plot of raw data over time. You can notice 48 distinct spikes that each correlate to an hour and half hour mark. This rounding off to the nearest approximate time value in the reports is why we chose to round off ALL values to the nearest hour.



- DATA VALUES Number of accidents rounded to each hour using round(time.count$Var1, units="hours") to round time to nearest hour.

Hour (A.M)	Accidents	Hour (P.M)	Accidents
12 A.M.	42279	12 P.M.	55999
1 A.M.	21430	1 P.M.	58868
2 A.M.	15877	2 P.M.	63423
3 A.M.	13028	3 P.M.	70400
4 A.M.	13078	4 P.M.	69964

5 A.M.	14213	5 P.M.	73928
6 A.M.	19022	6 P.M.	69163
7 A.M.	28924	7 P.M.	57342
8 A.M.	47870	8 P.M.	47682
9 A.M.	56182	9 P.M.	39668
10 A.M.	51595	10 P.M.	35399
11 A.M.	51317	11 P.M.	31924

Actual predicted accidents in each hour block:

Hour	**Predicted Accidents**	**Hour**	**Predicted Accidents**	**Hour**	**Predicted Accidents**
12 AM	29052.92	8 AM	48897.53	4 PM	68919.42
1 AM	19840.38	9 AM	57863.24	5 PM	72597.31
2 AM	15554.43	10 AM	52859.92	6 PM	68169.56
3 AM	12765.06	11 AM	52637.54	7 PM	56842.34
4 AM	12818.47	12 PM	56662.30	8 PM	47222.96
5 AM	17147.46	1 PM	59344.08	9 PM	38994.92
6 AM	21240.53	2 PM	63936.49	10 PM	34520.19
7 AM	30132.05	3 PM	71761.86	11 PM	30830.67

The Monte Carlo Error for the lambda values of each hour:

Hour	**MC Error**	**Hour**	**MC Error**	**Hour**	**MC Error**
12 AM	0.00075	8 AM	0.00451	4 PM	0.00672

1 AM	0.00177	9 AM	0.00655	5 PM	0.00738
2 AM	0.00229	10 AM	0.00529	6 PM	0.00658
3 AM	0.00401	11 AM	0.00763	7 PM	0.00507
4 AM	0.00394	12 PM	0.00578	8 PM	0.00522
5 AM	0.00494	1 PM	0.00514	9 PM	0.00672
6 AM	0.00453	2 PM	0.00578	10 PM	0.00794
7 AM	0.00377	3 PM	0.01079	11 PM	0.0091

- .bug File used in JAGS code

```
data {
}
model {
        for(i in 1:length(y)) {
                y[i] ~ dpois(lambda[i])
                log(lambda[i]) <- beta1 + beta2 * x[i]
        }
        beta1 ~ dnorm(0, 0.0001)
        beta2 ~ dnorm(0, 0.0001)
}
```

Everyone's contributions:

- David: Introduction
- Jefferson: Methods
- Aditya: Results & Conclusion
- Jake: JAGS and R coding, minor paper edits, and adding some Data to paper
- Everyone: data brainstorming, discussion of Bayesian methods, report editing and appendix additions