Analysis of Crash Frequency Using Hierarchical Negative Binomial Model with Bayesian Approach

Tingting Huang

INTRODUCTION

Crashes on urban interstates may cause huge economic losses and societal impacts due to the property damage, personal injury, travel delay and other issues. According to the estimation by National Highway Traffic Safety Administration [1], interstate highway crashes caused $28 billion in economic costs and $88 billion in comprehensive costs in 2010 in the U.S. In order to reduce crashes and improve traffic safety, it is necessary to identify the key factors that impact crash frequency on urban interstates.

From the traditional crash frequency studies, there are several kinds of factors which have impacts on the crash. The road geometry, traffic condition and weather are three main study objects in this project. To deal with the over-dispersed crash count data, this project employed a hierarchical negative binomial model with full Bayesian approach. This model was applied on the Interstate 235 (Des Moines, IA) dataset which can properly reflect the urban interstate situation. The significant factors were found and explained by interpreting the model estimation results. According to the results, the road geometry played an important role in monthly crash frequency.

DATA OVERVIEW

There were a total of 321 crashes on I-235 through lanes in 2013. Those crashes happened on different segments in different months. The data for modeling the crashes consisted of roadway characteristics, monthly traffic speed measures and monthly weather statistics.

The crash and roadway geometric information was obtained from Geographic Information Management System in Iowa Department of Transportation. In this system, every time the roadway geometric attributes (e.g. number of lanes) changed, the road was divided into segments at that point. Thus, the study road included 91 small directional segments with 1602 feet average length (shown in Figure 1).

The archived traffic speed data in 1-min interval were obtained from INRIX and traffic speed measures were computed at monthly aggregation level. Since the speed lower than 45 mph is generally considered as congestion on interstate, the percentage of time with average speed less than 45 mph in one month has been conducted to represent the congestion level.

Historical monthly weather data were downloaded from Quality Controlled Local Climatological Database. This project only extracted potentially effective weather variables like average temperature, average precipitation, maximum snow falls, etc.

After combining the data and removing the highly correlated variables, 6 variables could be used in this project. They are: segment length, number of lanes, left shoulder width, average speed, congestion index, maximum snow depth. The descriptive statistics of all the variables are shown in Table 1.

Additionally, the segment length is usually considered as an exposure variable to the crash count, which means the longer the segment is, the more crashes might happen on that segment.

Thus, in the modeling procedure, the natural logarithm has been taken on segment length data in order to treat it as the offset variable.

MODEL

The negative binomial model is derived from the Poisson model to handle the over-dispersion which is often to see in crash count data. Assuming the independence among crashes, the negative binomial regression model is given by:

$$Y_i \overset{ind}{\sim} NegBin(\lambda_i, \psi)$$

where Y_i is the i^{th} monthly crash count;

λ_i is the expectation;

ψ is the over-dispersion parameter;

$i = 1,2, \ldots, 1092$. (91 segments times 12 months)

The expectation λ_i is modeled as:

$$\lambda_i = \exp(\boldsymbol{X}_i{}^T \beta)$$

where $\boldsymbol{X}_i$ is the explanatory variables including ones as intercept;

β is the parameter vector.

To let the variables have random effect across segments, a hierarchical model is proposed:

$$\beta \overset{ind}{\sim} N(\mu_\beta, \Sigma_\beta)$$

where μ_β and Σ_β are the mean and covariance matrix for the multinomial normal distribution.

Here I assume ψ_i, μ_β and Σ_β are independent a priori. Thus the joint prior is:

$$p(\psi, \mu_\beta, \Sigma_\beta) = p(\psi)p(\mu_\beta)p(\Sigma_\beta)$$

Assigning the non-informative prior on those parameters, there are:

$$\psi \sim Gamma(1, 1)$$

$$p(\mu_\beta) \propto 1$$

For the covariance matrix Σ_β, instead of using the natural conjugate prior, the inverse-Wishart distribution, here I follows:

$$\Sigma_\beta = diag(\sigma_\beta)\Omega diag(\sigma_\beta)$$

where σ_β is a vector of standard deviations and Ω is a correlation matrix.

Now I have the standard deviation and correlation be independent a priori. Assigning priors on σ_β and Ω:

$$\sigma_\beta \sim Ca^+(0, 1)$$

$$p(\Omega) = |\Omega|^{\eta - 1}$$

η is the sharp parameter for the LKJ distribution, here assign $\eta = 1$.

RESULTS

To get the posterior distributions of the parameters, I ran 4 Markov chains in 1000 iterations (first 500 iterations as burn-in) with no thinning in Stan. The initial values and Metropolis-Hasting steps were set by Stan's default. Table 2 summarizes the information about the Markov chains. There is no evidence of non-convergence since the effective sample sizes are all greater than 100 and the absolute values of potential scale reduction factors are all less than 1.1. The traceplots (shown in Figure 2) of ψ and β are displaying a good mixture of chains which also indicates the adequate convergence of chains.

Table 2 also lists the mean, median and 95% credible interval for each parameter. In this project, β_1 is corresponding to the intercept and β_2 to β_7 are corresponding to those 6 variables. To better illustrate the credible interval, I put ψ and β_1 in one plot (Figure 3) since they have relatively larger scale than β_2 to β_7. The credible plot for β_2 to β_7 can be found in Figure 4, which has a smaller scale to accommodate the small values of those parameters.

In Figure 3 and Figure 4, the red bar indicates the 80% credible interval while the grey bar indicates the 95% credible interval. Also, the vertical black line is the zero line for easily illustrating if the parameter is statically significant or not. The β_1(intercept), β_6 (average speed) and β_7 (congestion level) is not significant at 80% credible level since the red bars involve zero. β_5 (maximum snow depth) is significant at 80% credible level but not at 95% level. Other parameters are all significant at 95% credible level. Those factors are the key factors we want to discuss in next section.

The posterior distribution for each parameter is shown in Figure 5. The distribution plot also includes the zero line to show how far or how close the majority of the distribution is toward the zero.

DISCUSSION

The over-dispersion parameter ψ is greater than 0 with mean value of 1.21, which indicates the data is over-dispersed. This is consistent with the situation shown in Table 1, where the monthly crash frequency's variance (0.65^2=0.4225) is greater than the mean (0.294). It validates the necessity of negative binomial model instead of Poisson model in this project. The discussion about all the critical variables are shown below.

Segment length (β_2): Since the crash frequency is supposed to increase on longer segment, segment length was treated as offset variable (a measure of exposure) and the natural logarithm of segment length was used in the model. The estimated mean of the parameter for log-segment -length is 0.69 which is a little lower than 1. That indicates the crash frequency is near linearly proportional to segment length, while with the increase of segment length, the increase rate in crash frequency is slightly lower.

Number of lanes (β_3): As number of lanes may account for more latent traffic flow variabilities, it may impact the crash frequency and was picked by the model. As a random parameter, the posterior is mainly positive, which indicates that crashes may increase as number of lanes increases. This result is consistent with some previous research. Milton [2] and Abdel-Aty [3] have found that number of lanes and crash rates have positive relations in their studies.

Left shoulder width (β_4): The left shoulder width was identified to have statistical significance. The estimated posterior is negative, which means the segments with wider left shoulder may have lower crash frequency. This is consistent with the previous findings about left shoulder width by Fitzpatrick [4]. In addition, it is also intuitively reasonable to assume that better lateral clearance helps reduce crashes.

Others: For weather information, the maximum snow depth (β_5) did not show the significant impact on crashes. However the most part of the posterior distribution was located at the right side of zero. This agreed with the intuition that one might have a higher chances experiencing crash during the heavier snow day. Regarding the traffic-related variables (β_6, β_7), there were no statistical evidence showing they had strictly positive or negative impact on crash frequency.

This project demonstrated an application of Bayesian hierarchical negative binomial model to analyze the key factors impacting monthly crash frequency on urban interstate. The model could better manage the over-dispersed data and capture the unobserved heterogeneity by allowing parameters to vary across segments with Bayesian approach. This exploratory study provided an insight of critical factors impacting crash frequency and explored the knowledge of urban interstate safety.

REFERENCES

[1] NHTSA. (2014). The Economic and Societal Impact of Motor Vehicle Crashes, 2010. National Highway Traffic Safety Administration, Publication DOT HS 812 013, Washington, D.C.

[2] Milton, J., and F. L. Mannering. (1998). The Relationship among Highway Geometries, Traffic-Related Elements and Motor-Vehicle Accident Frequencies. Transportation, Vol. 25(4), pp. 395–413.

[3] Abdel-Aty, M. A., and A. E. Radwan. (2000). Modeling Traffic Accident Occurrence and Involvement. Accident Analysis and Prevention, Vol. 32(5), pp. 633–642.

[4] Fitzpatrick, K., D. Lord, and B. Park. (2008). Accident Modification Factors for Medians on Freeways and Multilane Highways in Texas. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2083, pp. 62–71.

TABLES AND FIGURES

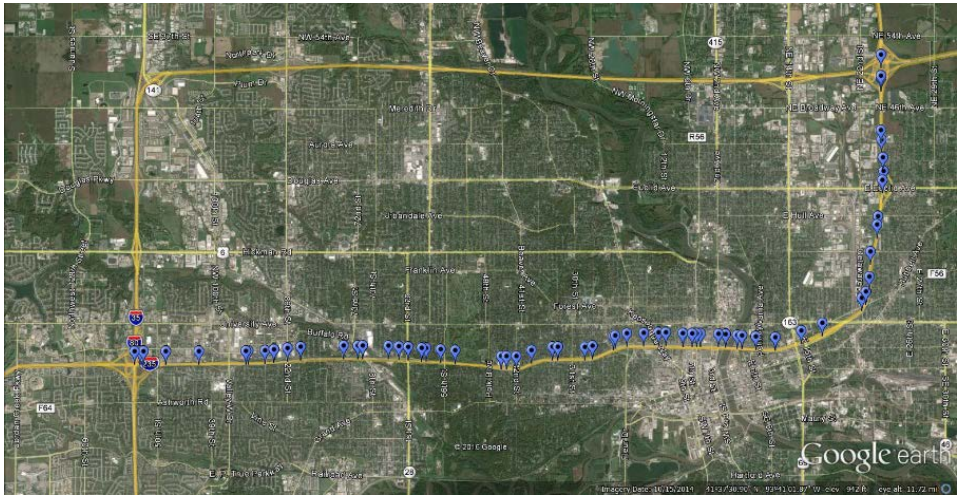

Figure 1 Locations of Segments in Study Area (blue pins)

Table 1 Descriptive Statistics of Variables

Variable	Mean	Std. Dev.	Min.	Max.
Dependent				
Monthly crash frequency	0.294	0.650	0.000	6.000
Independent				
Segment characteristics				
Segment length (mi)	0.303	0.225	0.008	0.980
Number of lanes	3.692	0.794	2.000	6.000
Left shoulder width (ft)	11.084	3.980	3.000	25.000
Traffic related information				
Monthly average speed (mph)	63.212	4.027	41.276	74.601
Percentage of time when speed lower than 45 mph (Congestion Index)	1.820	4.908	0.000	56.341
Weather information				
Maximum snow depth (in.)	3.000	3.561	0.000	9.000

Table 2 Information of Markov Chains for Each Parameter

Parameter	Corresponding Variables	Mean	2.5th Percentile	Median	97.5th Percentile	Effective Sample Size	$\hat{R}$
ψ	-	1.21	0.73	1.14	2.07	928	1.01
β_1	Intercept	-1.57	-4.05	-1.57	0.93	719	1
β_2	Length	0.69	0.52	0.69	0.87	978	1
β_3	Number of Lanes	0.33	0.15	0.33	0.51	1126	1
β_4	Left Shoulder Width	-0.06	-0.11	-0.06	-0.01	1172	1
β_5	Max. Snow Depth	0.03	-0.01	0.03	0.06	1188	1
β_6	Avg. Speed	0.01	-0.03	0.01	0.05	792	1
β_7	Congestion Level	0.01	-0.02	0.01	0.05	905	1
$lp\_\_$	-	-669.05	-680.75	-668.77	-659.38	139	1.01

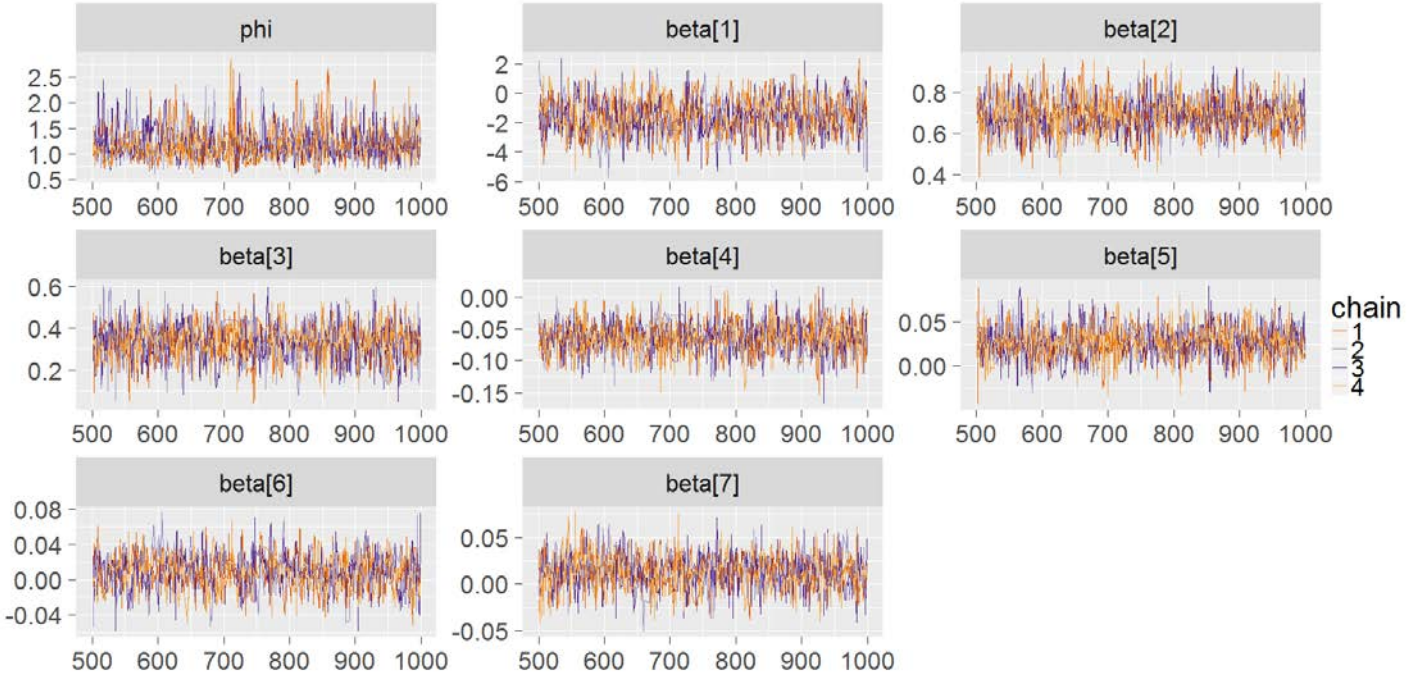

Figure 2 Traceplots of Four Chains without Warm-up for Each Parameter

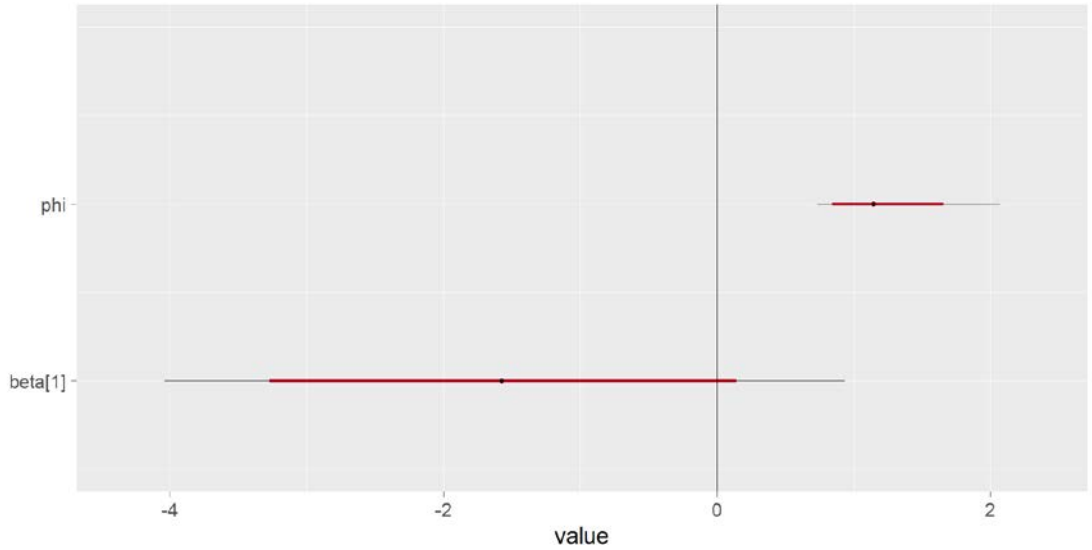

Figure 3 Credible Interval for ψ (over-dispersion) and β_1(intercept)

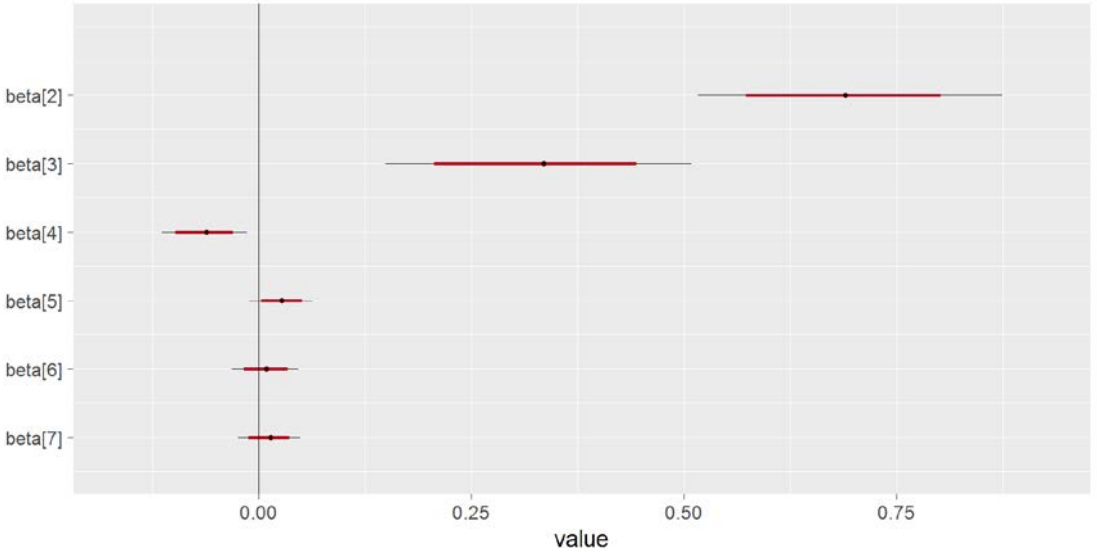

Figure 4 Credible Interval for β_2 to β_7

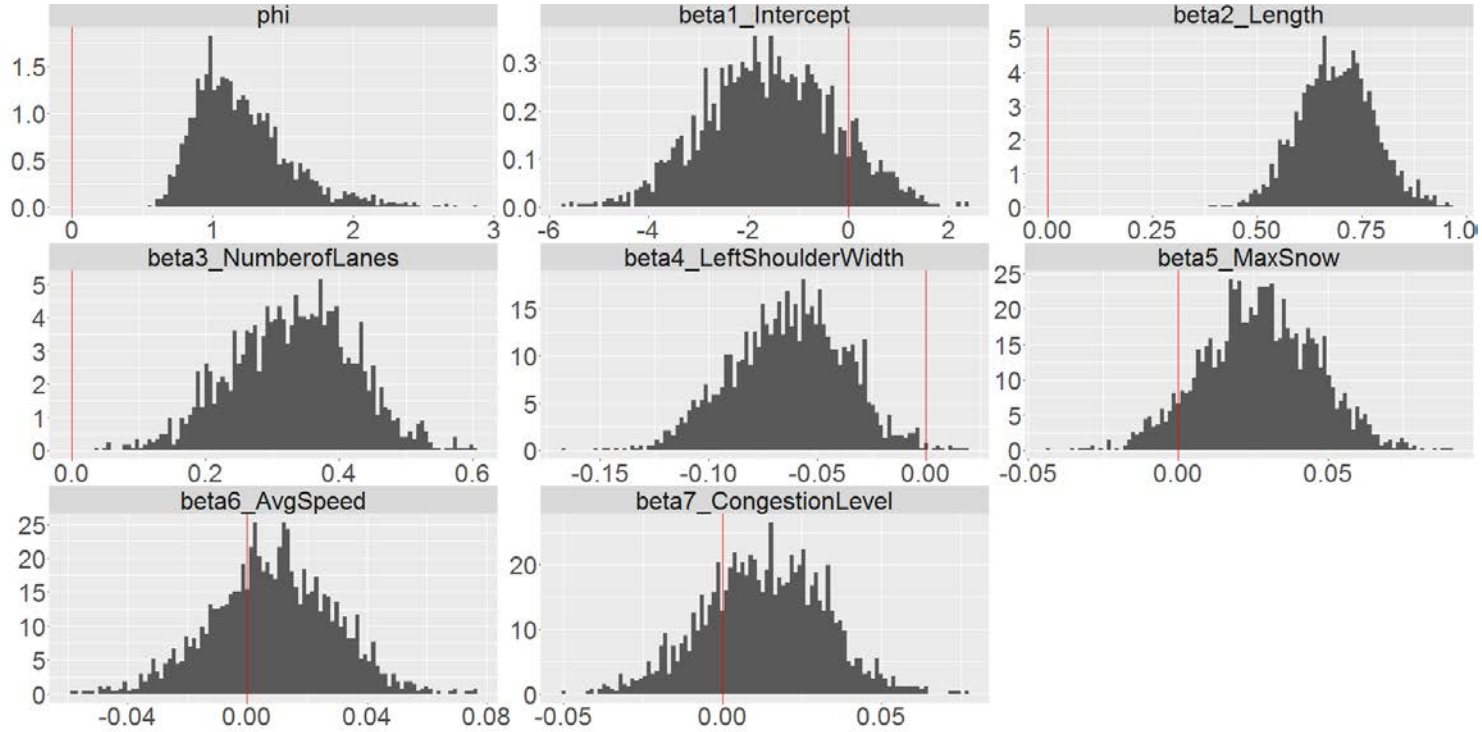

Figure 5 Posterior Distribution for Each Parameter