Bayesian Model for Vehicle Crashes on Two-Lane Primary Roads in Iowa

Daniel Ries
Michael D. Pawlovich
Alicia Carriquiry
Zachary Hans

April 25, 2017

Abstract

We explore the frequency and severity of crashes on two-lane primary roads in Iowa using a Bayesian generalized linear mixed effects model. We include models for all crashes, crashes resulting in fatalities and major injuries, crashes resulting in minor injuries, and crashed resulting in property damage only. Variables to be included in the models were chosen by a combination of Iowa Department of Transportation (DOT) employee expertise and the use of the spike and slab prior for variable selection in the `R` package `spikeSlabGAM`. Volume, district, and surface type were important for describing number of crashes for all models and federal function classification, urban area, and US or state road classification were important for all but the Fatal/Major model. Repeated measures random effects were included to induce correlation among years for given road segments. We used `Stan` to draw samples from the posterior distributions. We wrote software in `R` to perform model fitting as well as compute Safety Performance Functions (SPFs).

Contents

1 Introduction and Data Description

In this report we present a model as well as results to produce Safety Performance Functions (SPF) for two lane primary roads in Iowa. The model we develop and estimate uses data from 2005-2014. We have information on over 140 road characteristics such as speed limit, surface type, volume and access control (to name a few) for each road segment. We also have the total number of crashes, number of fatal crashes, number of crashes resulting in major injuries and minor injuries, and number of crashes resulting in only property damage. Because fatal crashes are relatively uncommon, we will combine fatal and major crashes into one category. There are 94,370 observations at 10,055 unique road segments.

2 Model Specification

To model the number of crashes as a function of observed covariates, we use a generalized linear mixed effects model with Poisson response and a log link. The random effect comes from the correlation induced between years at specific road segments. Because we have measurements at the same location over the course of 10 years, we include random effects to allow dependence across years. We also include an offset term as road segments are not all equal length.

Denote:

- Y_{ij}: number of crashes in segment i during year j
- t_i: length of segment i
- x_{ij}: p-dimensional vector of covariates for segment i and year j
- $\boldsymbol{\beta}$: p-dimensional parameter vector for covariate coefficients
- v_i: random intercept for segment i, induces correlation between years for segment i

The model is characterized by:

$$\begin{aligned} Y_{ij}|\lambda_{ij}, t_i &\overset{ind}{\sim} Poisson(\lambda_{ij} t_i), \\ log(\lambda_{ij})|\boldsymbol{\beta}, v_i &= x_{ij}^T \boldsymbol{\beta} + v_i, \\ v_i &\overset{iid}{\sim} N(0, \sigma_v^2). \end{aligned}$$

We specify the priors as:

$$\begin{aligned} \beta &\overset{iid}{\sim} N(0, 100), \\ \sigma_v^2 &\sim IG(0.1, 0.1). \end{aligned}$$

The term $\lambda_{ij} t_i$ is then the *expected number of crashes per mile* in segment i during year j.

3 Variable Selection

3.1 Spike and Slab Prior

There are over 140 road characteristics available, but we do not want to include all in the model in order to avoid overfitting. With the expertise of the Iowa Department of Transportation employees, the number of possible explanatory variables is narrowed down to about 40. However, many of these variables are factors that have many levels. We would like to narrow the possible variables down more so the models have a handful of interpretable variables. In addition, we don't want the factors that are included to have a large number of levels that could lead to overfitting and at the same time provide no real interpretation to the problem.

To select the possible explanatory variables to include in the model, we take a Bayesian approach by using spike and slab priors. We used the R package `spikeSlabGAM` to implement the variable selection. This package allows for variable selection in the generalized linear model case. It also can check for not only linear associations, but also higher order associations, interactions, as well as a general spline smoother. The article in the Journal of Statistical Software on this package is very helpful. The spike and slab prior that the authors of the `spikeSlabGAM` package in R developed is given by:

$$\begin{aligned} \beta|\gamma, \tau^2 &\sim N(0, v^2), \text{ where } v^2 = \tau^2\gamma, \\ \gamma|w &\sim wI_1(\gamma) + (1-w)I_{v_0}(\gamma), \\ \tau^2 &\sim Gamma(a_\tau, b_\tau), \\ w &\sim Beta(a_w, b_w). \end{aligned}$$

This amounts to a prior that puts a large amount of density close to 0 with some probability $1 - w$ for a regression coefficient β, and a dispersed density with some probability w. This provides shrinkage of the regression coefficient if it is close to 0, otherwise the coefficient is relatively unaffected by the prior. This setup allows us to get inclusion probabilities for each regression coefficient via γ. We included regression coefficients that had at least a 0.9 inclusion probability. After preliminary models were fit using the variables from the spike and slab prior, we utilized the expertise from DOT employees to do any necessary grouping of factors and eliminating variables that were included in all the models via variable selection, but appeared unimportant statistically and practically in the models. Table 1 gives the variables included in each model.

Model Fatal/Major	Model Minor	Model PDO	Model All Crashes
VOLUME	VOLUME	VOLUME	VOLUME
TRANSCENTE	TRANSCENTE	TRANSCENTE	TRANSCENTE
SURFTYPE	SURFTYPE	SURFTYPE	SURFTYPE
	FEDFUNC	FEDFUNC	FEDFUNC
	URBAN	URBAN	URBAN

Table 1: Variables included for each model.

3.2 Variable Descriptions

In this section we describe the variables that are present in the models as explanatory variables. For factor variables, we describe the levels we consider and any groupings. All groupings and levels are the same for the four models.

- VOLUME: describes usage of segment. We use log(VOLUME) as an explanatory variable.
- TRANSCENTE: factor indicating the district number. Takes values 1-6.
- SURFTYPE: Factor indicating surface type of the road. Grouped into two levels, asphalt (60,65,69,92) and concrete (70,74,76,77,79) as directed by DOT expertise. Numbers in () are from DOT Base Record Road and Structure Data document.
- FEDFUNC: Factor indicating the federal functional classification of the road segment. Grouped into two levels, other principal arterial (3) and minor arterial/major collector (4,5) as directed by DOT expertise.
- URBAN: Factor indicating whether road segment is in urban area or not.
- SYSCODE: Factor that indicates the state assigned system for the road segment. Takes two levels, US Route or Iowa Route.

4 Model Estimation

To estimate regression, variance, and random effect parameters of the models, we use MCMC to get draws from the posterior distributions. We used the program `Stan` in `R` via the package `rstan`. For each model, we ran 4 chains of length 10,000 with the first 5,000 used as burn-in. We ran the sampler in parallel so all 4 chains for a given model would sample at the same time. Running this on the hpc-class server took approximately 24 hours for each model. We checked trace plots and Gelman-Rubin diagnostics to assess convergence.

5 Results

5.1 Model for All Crashes

	Estimate	Posterior SD	2.5q	97.5q
(Intercept)	-8.694	0.392	-9.467	-7.925
TRANSCENTE2	-0.212	0.089	-0.385	-0.039
TRANSCENTE3	-0.025	0.085	-0.190	0.142
TRANSCENTE4	-0.037	0.088	-0.209	0.135
TRANSCENTE5	0.163	0.085	-0.004	0.327
TRANSCENTE6	0.113	0.092	-0.068	0.293
lVOLUME	0.760	0.051	0.659	0.861
FEDFUNC2	0.043	0.057	-0.068	0.154
URBAN	0.336	0.155	0.026	0.629
SURFTYPE2	-0.042	0.028	-0.098	0.013
sigmav	0.780	0.043	0.696	0.864

Table 2: Regression and variance parameter summaries for All Crashes Model

Notes on the fitted model:

1. TRANSCENTE: total crash rates appear to be lower in districts 2,3,4 while they are higher in districts 5 and 6 as compared to district 1.

2. lVOLUME: Volume is negatively associated with number of crashes. For everything else fixed, an increase in log(VOLUME) by 1, we'd expect a decrease in number of crashes by a factor of $e^{-0.317}$ =0.72.

3. FEDFUNC: The estimate of -0.496 for FEDFUNC2 means those roads classified as minor arterial or major collector (4,5) have an expected number of crashes $e^{-0.496}$ =60% of other principal arterial roads, all else equal.

4. SYSCODE3: The estimate of -0.275 for SYSCODE3 means Iowa Routes have an expected number of crashes $e^{-0.275}$ =76% of US Routes, all else equal.

5. URBAN: The estimate of 0.228 for URBAN means routes in designated urban areas have an expected number of crashes $e^{0.228}$ =125% of roads in rural areas, all else equal.

6. SURFTYPE2: The estimate of -0.077 for SURFTYPE2 means concrete roads have an expected number of crashes $e^{-0.077}$ =92% of asphalt roads, all else equal.

5.2 Model for Fatal and Major Injuries

	Estimate	Posterior SD	2.5q	97.5q
(Intercept)	-8.640	0.317	-9.266	-8.023
TRANSCENTE2	-0.224	0.089	-0.399	-0.050
TRANSCENTE3	-0.035	0.085	-0.200	0.132
TRANSCENTE4	-0.044	0.088	-0.218	0.126
TRANSCENTE5	0.152	0.084	-0.013	0.316
TRANSCENTE6	0.113	0.092	-0.070	0.291
lVOLUME	0.757	0.043	0.674	0.842
SURFTYPE2	-0.042	0.028	-0.097	0.013
sigmav	0.783	0.043	0.694	0.864

Table 3: Regression and variance parameter summaries for Fatal/Major injury Crashes Model

Notes on the fitted model:

1. TRANSCENTE: Fatal/major injury crash rates appear to be lower in zones 2,3,4 while they are higher in zones 5 and 6.

2. lVOLUME: Volume is negatively associated with number of crashes. For everything else fixed, an increase in log(VOLUME) by 1, we'd expect a decrease in number of crashes by a factor of $e^{-0.340}$ =0.71.

3. SURFTYPE2: The estimate of -0.091 for SURFTYPE2 means concrete roads have an expected number of crashes $e^{-0.091}$ =91% of asphalt roads, all else equal.

5.3 Model for Minor Injuries

	Estimate	Posterior SD	2.5q	97.5q
(Intercept)	-8.694	0.392	-9.467	-7.925
TRANSCENTE2	-0.212	0.089	-0.385	-0.039
TRANSCENTE3	-0.025	0.085	-0.190	0.142
TRANSCENTE4	-0.037	0.088	-0.209	0.135
TRANSCENTE5	0.163	0.085	-0.004	0.327
TRANSCENTE6	0.113	0.092	-0.068	0.293
lVOLUME	0.760	0.051	0.659	0.861
FEDFUNC2	0.043	0.057	-0.068	0.154
URBAN	0.336	0.155	0.026	0.629
SURFTYPE2	-0.042	0.028	-0.098	0.013
sigmav	0.780	0.043	0.696	0.864

Table 4: Regression and variance parameter summaries for Minimum injury Crashes Model

Notes on the fitted model:

1. TRANSCENTE: total crash rates for minor injuries appear to be lower in zones 2,3,4 while they are higher in zones 5 and 6.

2. lVOLUME: Volume is negatively associated with number of crashes. For everything else fixed, an increase in log(VOLUME) by 1, we'd expect a decrease in number of crashes by a factor of $e^{-0.093}$ =0.91.

3. FEDFUNC: The estimate of -0.615 for FEDFUNC2 means those roads classified as minor arterial or major collector (4,5) have an expected number of crashes $e^{-0.615}$ =54% of other principal arterial roads, all else equal.

4. SYSCODE3: The estimate of -0.119 for SYSCODE3 means Iowa Routes have an expected number of crashes $e^{-0.119}$ =89% of US Routes, all else equal.

5. URBAN: The estimate of 0.184 for URBAN means routes in designated urban areas have an expected number of crashes $e^{0.184}$ =120% of roads in rural areas, all else equal.

6. SURFTYPE2: The estimate of -0.064 for SURFTYPE2 means concrete roads have an expected number of crashes $e^{-0.064}$ =94% of asphalt roads, all else equal.

5.4 Model for Property Damage Only

	Estimate	Posterior SD	2.5q	97.5q
(Intercept)	-8.694	0.392	-9.467	-7.925
TRANSCENTE2	-0.212	0.089	-0.385	-0.039
TRANSCENTE3	-0.025	0.085	-0.190	0.142
TRANSCENTE4	-0.037	0.088	-0.209	0.135
TRANSCENTE5	0.163	0.085	-0.004	0.327
TRANSCENTE6	0.113	0.092	-0.068	0.293
lVOLUME	0.760	0.051	0.659	0.861
FEDFUNC2	0.043	0.057	-0.068	0.154
URBAN	0.336	0.155	0.026	0.629
SURFTYPE2	-0.042	0.028	-0.098	0.013
sigmav	0.780	0.043	0.696	0.864

Table 5: Regression and variance parameter summaries for PDO Crashes Model

Notes on the fitted model:

1. TRANSCENTE: total crash rates for PDO appear to be lower in zones 2,3,4 while they are higher in zones 5 and 6.

2. lVOLUME: Volume is negatively associated with number of crashes. For everything else fixed, an increase in log(VOLUME) by 1, we'd expect a decrease in number of crashes by a factor of $e^{-0.318}$ =0.73.

3. FEDFUNC: The estimate of -0.498 for FEDFUNC2 means those roads classified as minor arterial or major collector (4,5) have an expected number of crashes $e^{-0.498}$ =61% of other principal arterial roads, all else equal.

4. SYSCODE3: The estimate of -0.274 for SYSCODE3 means Iowa Routes have an expected number of crashes $e^{-0.274}$ =89% of US Routes, all else equal.

5. URBAN: The estimate of 0.228 for URBAN means routes in designated urban areas have an expected number of crashes $e^{0.228}$ =120% of roads in rural areas, all else equal.

6. SURFTYPE2: The estimate of -0.077 for SURFTYPE2 means concrete roads have an expected number of crashes $e^{-0.077}$ =94% of asphalt roads, all else equal.