Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)

Application of generalized estimating equations for crash frequency modeling with temporal correlation

Cite this as: Zhi-bin LI, Wei Wang, Pan LIU, Yong WANG, Li-teng ZHA, 2014. Application of generalized estimating equations for crash frequency modeling with temporal correlation. *Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)*, 15(7):529-539. [doi:10.1631/jzus.A1300342]

Background

Previously, numerous crash prediction models have been developed. Various methodologies have been proposed for crash frequency modeling to improve the predictive accuracy for crashes.

In most of the previous crash prediction models, crash counts were usually aggregated over several years and the average crash count per year was considered the response variable.

To enlarge the sample size in the dataset, a natural consideration is to divide the data aggregated over several years into smaller time intervals (a unit of year) and treat the crash counts in each year as separate observations.

However, disaggregating the crash data could create a temporal correlation in the dataset which could adversely affect the precision of parameter estimates in the crash prediction model if not properly considered.

The main objective is to evaluate the application of GEEs to account for the temporal correlation in the crash frequency modeling.

Data

The data were collected from 32 sections of exit ramps on the a freeway in China.

Four-year crash data, from 2006 to 2009, was obtained from the local freeway management agency. A total of 4429 crashes were observed.

The average crash count per year is 34.60 with a standard deviation (S.D.) of 38.87. The crash data have an obvious feature of over-dispersion.

There are two types of exit ramps that are typical on the Guangshen freeway, according to the number of exit lanes shown in Fig. 1.

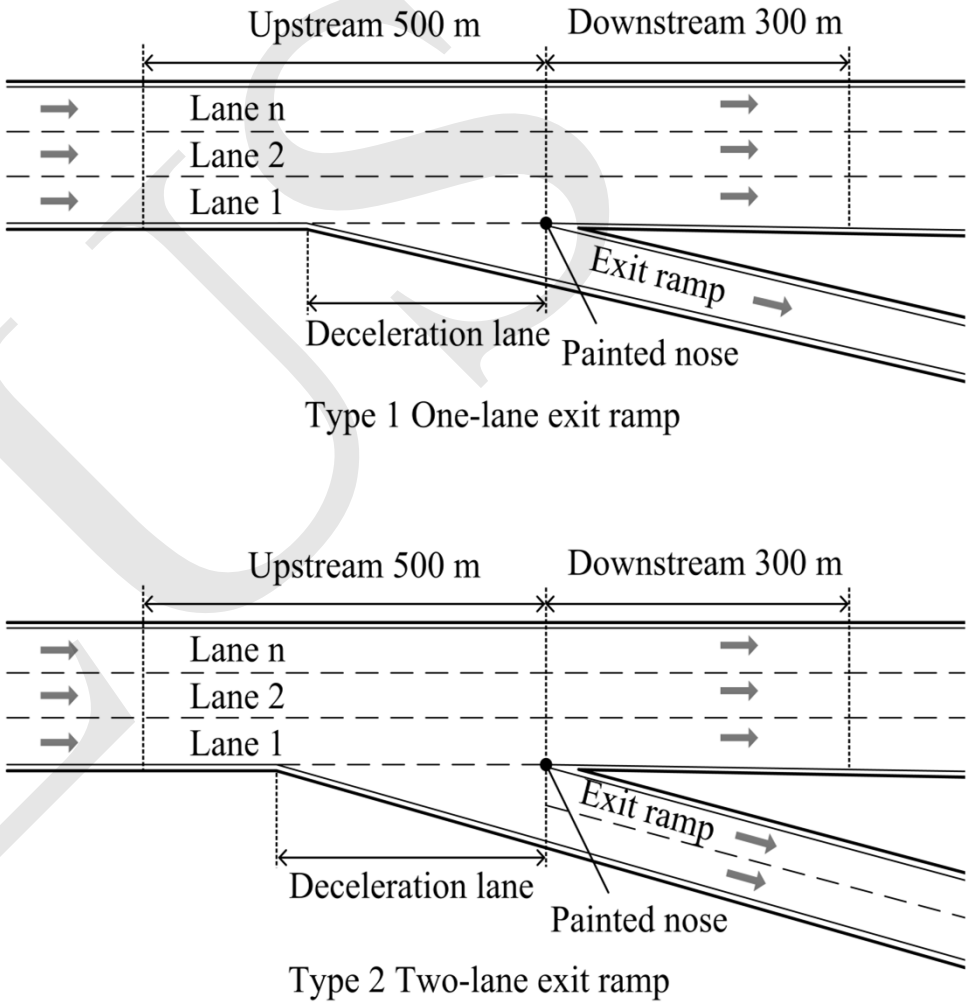


Fig. 1 Illustration of two types of exit ramps

Methodology

the model specifications for crash frequency analysis at freeway exit ramps in literature, the following model form is considered:

$$\ln(E\{u_t\})=\ln(\beta_0)+\beta_1\ln(F_1(t))+\beta_2\ln(F_2(t)) +\beta_3X_3(t)+\ldots+\beta_JX_J(t), \quad (1)$$

where $\ln(E\{\mu_t\})$ is the natural log of expected crash frequency in period t at the exit ramps, u_t is the crash frequency in period t, $F_1(t)$ and $F_2(t)$ are annual average daily traffic (AADT) on the mainline and ramps in period t, respectively, $X_j(t)$ is the jth explanatory variable in period t, and β_j is the jth coefficient to be estimated (j=0, 1, …, J), J is the number of coefficient.

The GEE is an extension of the GLM for estimating the temporally correlated data. Using the link function shown in Eq. (1), the coefficients β are estimated by

$$\sum_{i=1}^{I} \boldsymbol{D}_i^{\prime}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{u}_i)=0,$$

Results

Table 4 Model estimates of GLM with GEE procedure

Variable	Independent		Exchangeable		Autoregressive		Unstructured	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	−1.992	2.000	−1.667	2.259	−1.526	2.236	−1.862	2.309
Grade	0.323	0.189	0.377	0.207	0.402	0.209	0.445	0.182
Logarithm of AADT on mainline	0.520	0.229	0.470	0.251	0.465	0.252	0.530	0.250
Logarithm of AADT on ramp	0.244	0.103	0.266	0.110	0.220	0.108	0.181	0.118
Bad weather ratio	3.078	0.496	2.795	0.435	2.863	0.457	2.982	0.423
Right shoulder width	−0.112	0.072	−0.116	0.078	−0.098	0.076	−0.103	0.080
Summary statistics								
Cluster size	4		4		4		4	
Maximum absolute value	23.15		22.24		44.46		59.63	
P-value	0.470		0.618		0.092		0.026	

The GLM 2 has a better statistical fitness than the GLM 1 as shown in Table 3. The α means the level of dispersion in crash data. Smaller AIC (Akaike information criterion) and BIC (Bayesian information criterion) means the model fits the data better. These results show how the small sample size issue impacts the model estimates and leads to poor model performances.

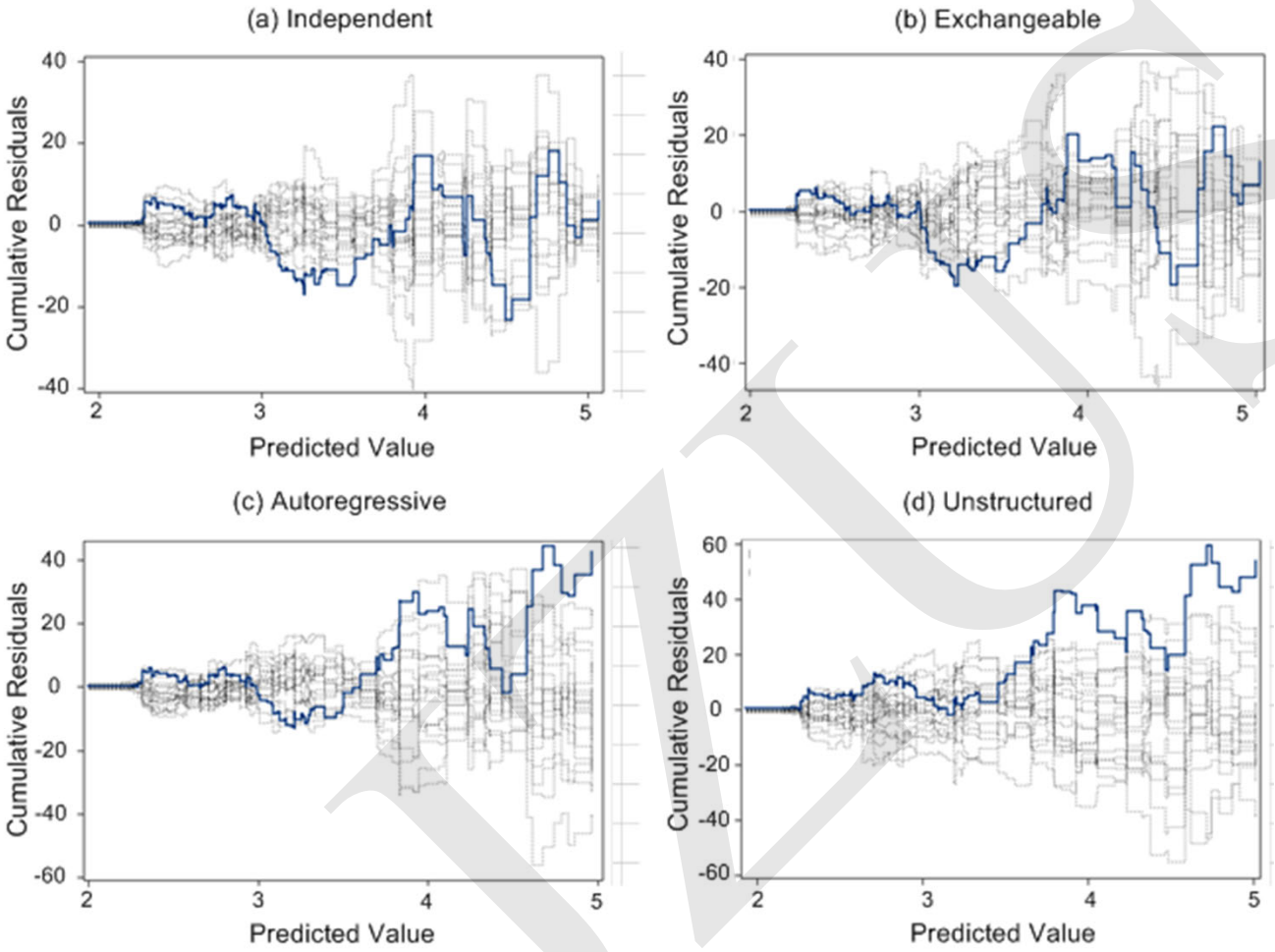
Results

Table 4 Model estimates of GLM with GEE procedure

Variable	Independent		Exchangeable		Autoregressive		Unstructured	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Intercept	−1.992	2.000	−1.667	2.259	−1.526	2.236	−1.862	2.309
Grade	0.323	0.189	0.377	0.207	0.402	0.209	0.445	0.182
Logarithm of mainline AADT	0.520	0.229	0.470	0.251	0.465	0.252	0.530	0.250
Logarithm of ramp AADT	0.244	0.103	0.266	0.110	0.220	0.108	0.181	0.118
Bad weather ratio	3.078	0.496	2.795	0.435	2.863	0.457	2.982	0.423
Right shoulder width	−0.112	0.072	−0.116	0.078	−0.098	0.076	−0.103	0.080
Summary Statistics								
Cluster size	4		4		4		4	
Maximum Absolute Value	23.15		22.24		44.46		59.63	
P-value	0.470		0.618		0.092		0.026	

the GLM 2 has a better statistical fitness than the GLM 1 as shown in Table 3. The α means the level of dispersion in crash data. Smaller AIC (Akaike information criterion) and BIC (Bayesian information criterion) means the model fits the data better. These results show how the small sample size issue impacts the model estimates and leads to poor model performances.

Three Representative Activities



The residuals for the GEE with exchangeable correlation structures are centered at zero and the plot of the residuals against any coordinate exhibits no systematic tendency.

Figure 3. Model assessments for GEEs with different correlation structures

Three Representative Activities

Table 6 Type III analyses for different models

Variable	GLM 2		GLM with GEE (Exchangeable)	
	Type III χ^2	*P*-value	Type III χ^2	*P*-value
Curve	4.97	0.0258	2.93	0.0868
Logarithm of mainline AADT	9.44	0.0021	3.04	0.0814
Logarithm of ramp AADT	10.30	0.0013	5.91	0.0151
Bad weather ratio	40.41	<0.0001	10.39	0.0013
Right shoulder width	3.22	0.0726	1.99	0.1579

As identified in Table 6, the type III χ^2 values in the GLM with GEE are generally smaller than that in the GLM 2 and the *P*-values for variables are larger in the GLM with GEE. It indicates that the traditional GLM without accounting for the temporal correlation would overestimate the significance of predicting factors.

As shown in Table 6, the right shoulder width is estimated to be significantly related to crash counts at a 90% confidence level in the GLM 2. But after considering the temporal correlation in the GLM with GEE, this variable becomes insignificant at the same confidence level.

Conclusions

This study evaluated the application of the GEE to account for the temporal correlation in the crash frequency data. Using four-year crash data at exit ramps on a freeway, the GLM with GEE was estimated based on yearly disaggregated crash data. For comparison purposes, traditional GLMs were also estimated based on the same dataset.

The results showed that there were significant temporal correlations in the yearly disaggregated crash data used in this study. The GEE procedure captured the correlation among crash counts in different years. The exchangeable correlation structure fitted the data properly. A comparison between the GLM and the GLM with GEE showed that the traditional GLM could underestimate the standard errors of explanatory variables and make incorrect inferences on the significance of the variables. The GLM with GEE captured the features of temporal correlation in the data and led to more accurate estimates on the impacts of predictors.