Journal of Statistical Research *https://doi.org/10.47302/jsr.2021550111*
2021, Vol. 55, No. 1, pp. 159-174 *ISSN 0256 - 422 X*

A FLEXIBLE GEV LINK FOR ZERO INFLATED CONWAY MAXWELL POISSON REGRESSION WITH A CASE STUDY ON MOTOR VEHICLE CRASHES

XIAOMENG LI

Department of Statistics, University of Connecticut, Connecticut, USA
Email: xiaomeng.li@uconn.edu

DIPAK K. DEY*

Department of Statistics, University of Connecticut, Connecticut, USA
Email: dipak.dey@uconn.edu

SUMMARY

This paper introduces a new flexible link for zero inflated Conway Maxwell Poisson (COM-Poisson) distribution. Zero inflated Poisson regression has been widely used for modeling rare events with excess zeros. In recent years, the zero inflated Conway Maxwell Poisson regression has been proposed. The advantage of COM-Poisson is its ability to handle both under- and over-dispersion through controlling one special parameter in the distribution, which makes it more flexible than current frequently used models, i.e., Poisson and Negative Binomial. The usual link function for zero inflated models is the logit link, which assumes the response curve between covariates and the probability of zeros is symmetric. This assumption is not always true. To add more flexibility, we propose a new flexible link function for the zero inflated Conway Maxwell Poisson regression, the generalized extreme value (GEV) model, which can capture different skewness with a shape parameter. Thus we can let data tell the skewness of the link function. Simulation studies and an application on traffic accident data are conducted to show the flexibility of our proposed model against the commonly used models.

Keywords and phrases: COM-Poisson model, Bayesian inference, GEV link, Model Comparison, Zero Inflated

1 Introduction

Regression is the commonly used statistical method to explore the relationship between the variable of interest and predictors. The variable of interest is count data in many applications, and the commonly used model for count response variable is Poisson regression. To accommodate the mean variance equivalence assumption of Poisson distribution, Negative Binomial regression is often adopted for over-dispersed data. Generalized Poisson regression (Famoye, 1993) can model both

* Corresponding author

over- and under-dispersion. However, it belongs to exponential family only for a constant dispersion parameter (Cui et al., 2006). The dispersion parameter of generalized Poisson model cannot be linked with predictors. A more general distribution for count data is Conway Maxwell Poisson distribution (Shmueli et al., 2005), which can also handle both over- and under-dispersed data by introducing a dispersion parameter. A dual-link generalized linear model based on Conway Maxwell Poisson is well developed (Guikema and Goffelt, 2008; Sellers and Shmueli, 2010). Both mean and dispersion parameter can be modelled by predictors. Thus in this paper, we focus on Conway Maxwell Poisson model.

As for rare events like disease and accident, there are usually excess zeros. Zero inflated model, which is a mix of degenerate distribution at zero with count distributions, is a common way to deal with excess zeros. In most cases, zero inflated Poisson model is considered (Lambert, 1992). For additional over-dispersion, zero inflated Negative Binomial model is adopted (Ridout et al., 2001). Recently zero inflated Conway Maxwell Poisson model has been proposed (Barriga and Louzada, 2014; Sellers and Raim, 2016). The usual link function for zero inflated models is the logit link, which is a symmetric link. It assumes that the probability of zeros approaches to 0 at the same speed as approaching to 1. This assumption is not always true. A wrong choice of link function can result in poor fit of the model and the importance of choosing the appropriate link function has been shown in many previous studies (Nagler, 1994; Chen et al., 1999). Whether the link function is symmetric or skewed and in which direction it is skewed is often unknown to us. Thus in this paper, we propose the flexible generalized extreme value (GEV) link (Wang and Dey, 2010, 2011), which has a shape parameter to capture the skewness, to the zero inflated Conway Maxwell Poisson model. The shape parameter can be estimated by data, thus we can let data tell the skewness of the link function.

Considering traffic accident is rare event and usually exhibits excess zeros, we use traffic accident data as an application of proposed zero inflated Conway Maxwell Poisson model with GEV link. Poisson and Negative Binomial regression have been commonly applied on traffic accident data (Fridstrøm et al., 1995; Poch and Mannering, 1996). Aguero-Valverde compared Negative Binomial, Poisson Lognormal, zero inflated Poisson, zero inflated Negative Binomial and zero inflated Poisson Lognormal regressions with and without temporal effect and found Negative Binomial model with fixed over time random effects has best model fit (Aguero-Valverde, 2013). Generalized Poisson regression has been applied on traffic accident data and shown superiority than Poisson regression (Famoy et al., 2004). Recently, Conway Maxwell Poisson model has also been applied on traffic accident (Lord et al., 2008) and performs as good as Negative Binomial model. Zero inflated Conway Maxwell Poisson model has not been applied on traffic accident data yet.

This paper is organized as follows. In Section 2, we will review the Conway Maxwell Poisson distribution and the zero inflated Conway Maxwell Poisson model with GEV link. Simulation studies are shown in Section 3. In simulation studies, we first compare zero inflated Poisson model with GEV link with zero inflated Poisson model with logit link. We then compare zero inflated Conway Maxwell Poisson model with zero inflated Poisson model using GEV link under different scenarios. In Section 4, we show the application on traffic accident data, including data description, estimated parameters and model comparisons. Conclusions are given in Section 5.

2 Methodology

2.1 COM-Poisson distribution

Conway Maxwell Poisson (COM-Poisson) distribution, which was first introduced in 1962 for modeling queues and service rates by Conway and Maxwell (1962), has recently been re-introduced by statisticians to model count data. In this paper, we use the re-parameterization of the COM-Poisson distribution proposed by Guikema and Goffelt (2008), which provides good basis for developing generalized linear model. The probability mass function is shown in equation 2.1.

$$P(Y = y \,|\, \mu, \nu) = \Big(\frac{\mu^y}{y!}\Big)^{\nu} \frac{1}{Z(\mu, \nu)}, \tag{2.1}$$

where μ is mean parameter, ν is dispersion parameter, and $Z(\mu, \nu) = \sum_{j=0}^{\infty} (\mu^j / j!)^{\nu}$ is the normalization constant, which is analytically intractable. In this paper, we use the truncation method to approximate the normalization constant. Shmueli et al. (2005) derived the approximation of mean and variance of COM-Poisson distribution by approximating $Z(\mu, \nu)$ using asymptotic expression. By modifying that, the approximated mean and variance for the re-parameterized COM-Poisson distribution are shown in equation 2.2 and the mode of the COM-Poisson distribution is $\lfloor \mu \rfloor$.

$$\begin{aligned} E(Y) &\approx \mu + \frac{1}{2\nu} - \frac{1}{2}, \\ Var(Y) &\approx \frac{\mu}{\nu}. \end{aligned} \tag{2.2}$$

The dispersion parameter ν makes COM-Poisson distribution more flexible than Poisson distribution. When $\nu < 1$, COM-Poisson can model the over-dispersion of count data. While $\nu > 1$, it can capture the under-dispersion of count data. In this set up, μ closely approximate the mean, unless μ or ν is small. Both mean and dispersion parameters can be linked with predictors as shown in equation 2.3. For $i = 1, 2, \ldots, n$,

$$\begin{aligned} \log \mu_i &= \mathbf{X_i}'\beta, \\ \log \nu_i &= -\mathbf{X_i}'\delta. \end{aligned} \tag{2.3}$$

Then $E(Y_i) \approx \exp\{x_i'\beta\}$ and $V(Y_i) \approx \exp\{x_i'\beta + x_i'\delta\}$. Both mean and variance are allowed to vary for different values of predictors.

2.2 Zero inflated Conway Maxwell Poisson model

To capture excess zeros, a new parameter p is added in zero inflated model, where p is the probability of excess zeros and $(1 - p)$ is the probability of the count generating from a COM-Poisson distribution. The probability mass function of zero inflated COM-Poisson is in equation 2.4. For $i = 1, 2, \ldots, n$,

$$P(Y_i = y_i | \mu_i, p_i) = \begin{cases} p_i + \frac{1-p_i}{z(\mu_i, \nu_i)} & \text{when } y_i = 0, \\ (1 - p_i)\Big(\frac{\mu_i^y}{y_i!}\Big)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)} & \text{when } y_i > 0. \end{cases} \tag{2.4}$$

The parameter p is modeled as $p_i = F(x_i'\alpha)$ where F is a cumulative distribution function, and the inverse function of F determines the link function, that is $F^{-1}(p_i) = X_i'\alpha$. The usual link for p is the logit link, which is $F^{-1}(p_i) = \log\{p_i/(1-p_i)\}$. In this paper, we use a more flexible link function, the generalized extreme value (GEV) link to fit the probability p of zeros. GEV link is proposed based on the GEV distribution (Wang and Dey, 2010, 2011). The cumulative distribution function of GEV distribution is shown in equation 2.5,

$$G(x) = \exp\left[-\left\{1 + \xi\frac{(x-\mu)}{\sigma}\right\}_+^{-1/\xi}\right], \tag{2.5}$$

where $\xi \in R$ is the shape parameter controlling the tail behavior of the distribution and $x_+ = \max(x, 0)$. When $\xi \to 0$, $G(x) = \exp[-\exp\{-(x-\mu)/\sigma\}]$, and it gives the Gumble distribution. The GEV link is obtained by setting the F as the GEV distribution with $\mu = 0$ and $\sigma = 1$ as equation 2.6, assuming Z_i is a binary variable,

$$p_i = P(Z_i = 1) = 1 - \exp\left\{ -(1 - \xi x_i'\alpha)_+^{-1/\xi}\right\} = 1 - GEV(-x_i'\alpha; \xi). \tag{2.6}$$

The flexibility of GEV link comes from its ability to fit the skewness in the response curve with the free shape parameter ξ. When $\xi < \ln 2 - 1$, the GEV link model is negatively skewed, and when $\xi > \ln 2 - 1$, it is positively skewed. The range of skewness provided by GEV link is also much wider than commonly used skewed Cloglog link and the skewed generalized t-distribution link. Symmetric link is a special case which can be approximated by the class of GEV links.

2.3 Model comparison criteria

We use mean absolute error (MAE), Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) and Log-pseudo marginal likelihood (LPML) (Ibrahim et al., 2001) to measure the model performance. MAE is the mean absolute value of the difference between fitted value and the observations. DIC measures how well the model fits the data and also penalizes the complexity of model. It is defined in equation 2.7

$$DIC = \bar{D} + p_D, \tag{2.7}$$

where $D = -2\log L(\theta|y)$, which is the deviance. $\bar{D} = E(D)$ is the posterior mean of deviance, and p_D is the effective number of parameters, and is given by $p_D = \bar{D} - D(\bar{\theta})$, where $D(\bar{\theta})$ is a point estimate of the deviance obtained by plugging in the posterior mean of θ. Preferred model will have lower DIC.

LPML is used to measure the predictive ability of the model, which is the sum of log conditional predictive ordinate (CPO). CPO is based on leave one out cross validation, which estimates the probability of y_i when observing y_{-i}. y_{-i} is our sample when leave y_i out. Definition of CPO is shown in equation 2.8 and LPML is shown in equation 2.9. Preferred model will have higher LPML.

$$\begin{aligned} CPO_i &= f(y_i|y_{-i}) \\ &= \left[\int \frac{1}{f(y_i|\theta)} f(\theta|y) d\theta\right]^{-1} \end{aligned} \tag{2.8}$$

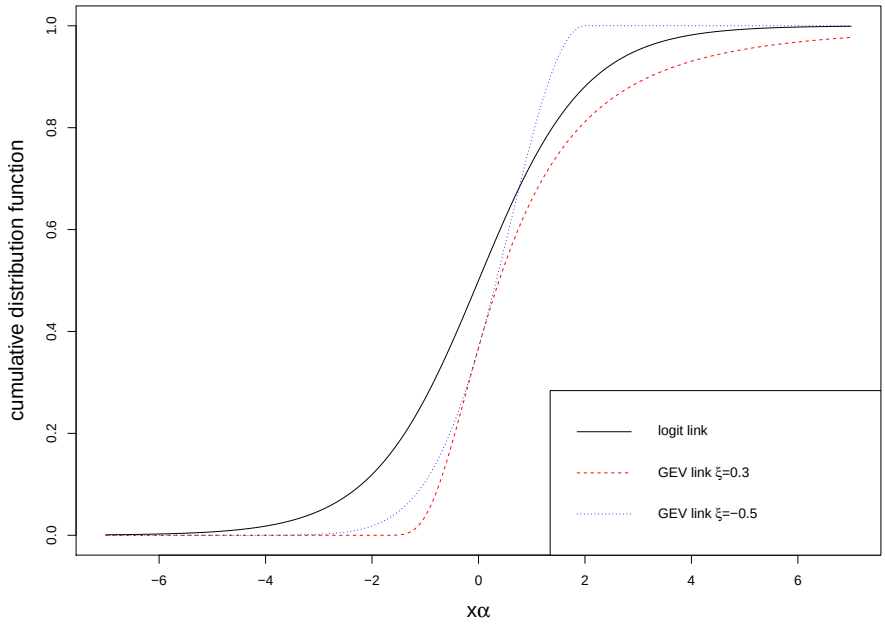

Figure 1: cumulative distribution functions of link functions with different skewness

$$LPML = \frac{1}{n}\sum_{i=1}^{n}\log(CPO_i). \tag{2.9}$$

3 Simulation

Model performances of zero inflated Poisson model with logit link and with GEV link are compared using simulated data under three scenarios. We simulate data following zero inflated Poisson model, with symmetric, positive skewed and negative skewed link functions. We use logit link as the symmetric link. As for skewed link, we use GEV link with $\xi = 0.3$ as positive skewed link and with $\xi = -0.5$ as negative skewed link. The cumulative distribution function of the three link functions that we use to simulate data are shown in the Figure 1. Black line represents logit link. Blue line is the GEV link with $\xi = -0.5$, which is negative skewed and red line is the GEV link with $\xi = 0.3$, which is positive skewed.

For each scenario, 5000 observations are generated. For each observation $(i = 1, 2, \ldots, 5000)$, first we generate a binary variable with probability $p_i = F(X_i'\alpha)$ being 1, where F is the corresponding cumulative distribution of the link function, $\alpha = (\alpha_0, \alpha_1)$ is the coefficient and $X = (X_0, X_1)$ is the covariate. $X_0 = 1$ is the intercept and set to be 1. X_1 is generated following normal

distribution with mean 1 and standard deviation 2. If this binary variable is 1, then we generate the observation $y_i = 0$. If the binary variable is 0, then we generate the observation following a Poisson distribution with mean parameter $\lambda_i = X_i'\beta$, where $\beta = (\beta_0, \beta_1)$ is the coefficient. The coefficients for link function we set $\alpha_0 = -1$ and $\alpha_1 = 0.5$, and the coefficients for mean parameter we set $\beta_0 = 1$ and $\beta_1 = 0.5$. For each scenario, we conduct the simulation on 50 dataset generated from the same model, and show the estimated coefficients in the Table 1. The coefficients are estimated under Bayesian framework. Non informative Normal priors are used for all the parameters. Metropolis-Hasting algorithm and Gibbs sampler are used to sample parameters from their posterior distributions. First 5000 iterations were discarded as burnin and the following 15000 iterations were collected as samples to study the posterior distributions of parameters. Traceplots, Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) and Geweke's diagnostics (Geweke, 1992) were used to measure the convergence of the Gibbs sampler. High performance computing (HPC) and parallel computation are adopted to save running time. The three criteria that we use to measure model fitting are shown in Table 2.

Table 1: Coefficient Estimates

		$\hat{\beta}_0$ (C_{95})	$\hat{\beta}_1$ (C_{95})	$\hat{\alpha}_0$ (C_{95})	$\hat{\alpha}_1$ (C_{95})
symmetric	logit	0.982 (0.880)	0.489 (0.920)	-0.983 (0.960)	0.492 (0.960)
	gev	0.982 (0.890)	0.490 (0.910)	-0.971 (.930)	0.310 (.000)
positively skewed	logit	0.972 (0.540)	0.515 (0.520)	-0.797 (0.080)	0.761 (.000)
	gev	0.999 (0.940)	0.502 (0.940)	-1.005 (.953)	0.501 (0.953)
negatively skewed	logit	1.013 (0.807)	0.495 (0.853)	-1.235 (.000)	0.875 (.000)
	gev	1.001 (0.953)	0.500 (0.967)	-1.004 (0.967)	0.500 (0.947)

Note: The C_{95} shows the coverage rate that the 95% credible interval of the parameter covers the true value that we set.

Table 2: Calculated Model Comparison Criteria

	symmetric		positively skewed		negatively skewed	
	logit	gev	logit	gev	logit	gev
MAE	3.237	3.231	2.229	2.118	2.398	2.400
DIC	16751.52	16751.66	13718.72	13666.81	14700.4	14692.26
LPML	-1.68	-1.69	-1.37	-1.37	-1.47	-1.47

"Symmetric", "positively skewed" and "negatively skewed" represent the three link function we use to generate the data. "logit" and "gev" represent the link function we use in the zero inflated Poisson model to fit the simulated data. As for parameter estimates, our comparison focus on β. If the true link function is different from the link function that we use, we do not expect the estimated α being the same as the α that we set. From Table 1 we can see the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ and the coverage rate of them using two links are close to each other. The $\hat{\alpha}_1$ of model using GEV link is different from the true value we set as we expect. When the true link function is positively skewed or negatively skewed, we can find the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ when using model with logit link are not that close to true value. Especially for the positively skewed true link scenario, the coverage rate of β_0 and β_1 are just above 0.5. Thus if we use wrong link function, the mean coefficient estimate will also be influenced. The $\hat{\alpha}_0$ of model using logit link is smaller than the true value and the $\hat{\alpha}_1$ is greater than the true value if the true link function is not symmetric. The estimated coefficients using model with GEV link are all very close to the true value that we set.

From Table 2 we can find LPML of models using logit link and using GEV link perform similar. Based on DIC, GEV link performs better than logit link. GEV link improves MAE slightly when comparing with logit link for some scenarios, but the MAE of both models are small and close to each other. Overall speaking, zero inflated Poisson with GEV link performs as good as logit link when the true link function is logit and performs better than logit link when the true link is skewed based on this simulation study.

We also perform another simulation study to compare the zero inflated Conway Maxwell Poisson model with the zero inflated Poisson model under 9 different scenarios, the combination of over-, equi- and under-dispersion of data with positive, negative and symmetric link function of probability of zeros. Both zero inflated Conway Maxwell Poisson model and zero inflated Poisson model are using GEV link for this study. Due to the longer running time of zero inflated Conway Maxwell Poisson model, we generate 2000 observations for each dataset and conduct simulation on 10 dataset generated from the same model for each scenario. As for the positive, negative and symmetric link function, we use the same parameter settings as the previous one. For equi-dispersed scenario, we generate data from zero inflated Poisson distribution. For over- and under-dispersed scenario, we generate data from zero inflated Conway Maxwell Poisson with different dispersion parameter ν. The dispersion parameter is linked with predictors with a log link as introduced in methodology section, that is $\log \nu_i = -\mathbf{X_i}'\delta$, where $\delta = (\delta_0, \delta_1)$. We set $\delta_1 = 0$, $\delta_0 = -0.5$ for over-dispersed data and $\delta_0 = 1$ for under dispersed data. The data generation steps are the same. We first generate the binary variable with probability p linked with predictors using three link functions. For equi-dispersed scenario, when the binary variable is 0, we generate the observation following Poisson distribution. As for over- and under-dispersed scenarios, we generate the observation following Conway Maxwell Poisson distribution. When generating observations following Conway Maxwell Poisson distribution, we use the rejection sampling method (Chanialidis et al., 2018). The simulation results are shown in the Table 3 and Table 4.

Table 3: Coefficient Estimates

dispersion	link	model	$\hat{\beta}_0$ (C_{95})	$\hat{\beta}_1$ (C_{95})	$\hat{\alpha}_0$ (C_{95})	$\hat{\alpha}_1$ (C_{95})	ξ (C_{95})
equi-	symmetric	ZIP	1.013 (0.9)	0.495 (0.9)	-1.011 (1)	0.316 (0)	-0.230
		ZICMP	1.012 (0.9)	0.496 (0.9)	-1.010 (1)	0.315 (0)	-0.237
	positively skewed	ZIP	1.013 (0.85)	0.494 (0.85)	-1.029 (0.95)	0.502 (1)	0.395 (0.9)
		ZICMP	1.005 (0.95)	0.496 (0.85)	-1.032 (.95)	0.502 (1)	0.398 (0.8)
	negatively skewed	ZIP	1.011 (0.9)	0.492 (0.8)	-0.982 (1)	0.490 (1)	-0.522 (0.95)
		ZICMP	1.009 (1)	0.493 (1)	-0.983 (1)	0.490 (1)	-0.520 (0.9)
over-	symmetric	ZIP	1.380 (0)	0.401 (0)	-0.920 (0.95)	0.284 (0)	-0.047
		ZICMP	1.006 (0.95)	0.493 (0.85)	-0.984 (.9)	0.313 (0)	-0.286
	positively skewed	ZIP	1.417 (0)	0.369 (0)	-0.806 (0.15)	0.413 (0)	0.609 (0.5)
		ZICMP	0.970 (0.9)	0.516 (0.85)	-0.973 (1)	0.490 (0.95)	0.297 (1)
	negatively skewed	ZIP	1.420 (0)	0.373 (0)	-0.944 (0.9)	0.463 (0.6)	-0.154 (0)
		ZICMP	1.082 (0.75)	0.466 (0.8)	-1.010 (1)	0.501 (1)	-0.401 (0.95)
under-	symmetric	ZIP	0.875 (0)	0.535 (0)	-0.963 (0.9)	0.315 (0)	-0.498
		ZICMP	0.996 (0.9)	0.504 (0.9)	-0.947 (.8)	0.305 (0)	-0.332
	positively skewed	ZIP	0.860 (0)	0.553 (0)	-1.096 (0.8)	0.545 (0.7)	0.160 (0.9)
		ZICMP	1.005 (1)	0.501 (1)	-1.014 (.8)	0.497 (0.8)	0.425 (0.95)
	negatively skewed	ZIP	0.863 (0)	0.545 (0)	-1.017 (0.9)	0.518 (0.9)	-0.648 (0.7)
		ZICMP	1.003 (0.9)	0.499 (0.8)	-1.016 (0.8)	0.500 (0.9)	-0.446 (1)

Table 4: Calculated Model Comparison Criteria

		symmetric		positively skewed		negatively skewed	
dispersion		ZIP	ZICMP	ZIP	ZICMP	ZIP	ZICMP
equi-	MAE	3.158	3.402	2.021	1.962	2.191	2.020
	DIC	6761.777	6764.828	5506.459	5502.06	5897.797	5901.16
	LPML	-1.706	-1.708	-1.410	-1.415	-1.509	-1.510
over-	MAE	2.206	2.416	1.518	1.236	1.536	1.883
	DIC	8140.254	7750.584	6543.176	6355.126	7048.478	6809.183
	LPML	-2.049	-2.030	-1.723	-1.611	-1.769	-1.723
under-	MAE	1.826	2.629	1.149	0.981	2.072	2.097
	DIC	6809.81	6196.5	5528.427	5132.184	5955.469	5404.814
	LPML	-1.604	-1.56	-1.294	-1.318	-1.413	-1.379

In Table 4, "equi-","over-","under-" represent the scenarios when the generated data are equi-, over- and under-dispersed. "ZIP" means the model we use to fit the data is zero inflated Poisson and "ZICMP" means the model we use to fit the data is zero inflated Conway Maxwell Poisson. As for equi-dispersed data, the estimated coefficients of both models are close to the true parameters that we set, except the $\hat{\alpha}_1$ and ξ when the true link is symmetric, which is what we expect. As for over- and under- dispersed version, the most estimated coefficients of zero inflated Poisson model are different from what we set. When we take a look at the model comparison criteria in Table 4, zero inflated Conway Maxwell Poisson model improves the MAE for some scenarios. The MAE of both models are small and close to each other. For equi-dispersed scenario, DIC and LPML of both models are close to each other. Zero inflated Conway Maxwell Poisson model performs as good as zero inflated Poisson model when the data is equi-dispersed. As for over- and under- dispersed situation, we can see an obvious drop of DIC and increase of LPML when we use zero inflated Conway Maxwell Poisson model. From the simulation results, we can see the flexibility of zero inflated Conway Maxwell Poisson model and it can handle different scenarios quite well.

4 Real Data Analysis

We then applied the zero inflated Poisson and zero inflated Conway Maxwell Poisson model with GEV link to the infrastructure safety evaluation dataset. Vehicle accidents are rare events, which are commonly modeled by zero inflated Poisson. In this study, we would like to show that the zero inflated Conway Maxwell Poisson model performs better than the commonly used zero inflated Poisson model. Suppose Y_i represents the number of crashes for ith observation. We assume $Y_i \sim ZIP(\mu_i E_i, p_i)$ for zero inflated Poisson model and $Y_i \sim ZICMP(\mu_i E_i, \nu_i, p_i)$ for zero inflated

Conway Maxwell Poisson model where μ_i is the mean parameter, E_i is the exposure, ν_i is the dispersion parameter and p_i is the probability of zeros. Huang (2017) shows that exposures can be taken into account by including offset in the Conway Maxwell model. μ_i, ν_i and p_i are modeled by covariates as shown in methodology parts, that is

$$\begin{aligned} \log \mu_i &= \mathbf{X_i}'\beta, \\ \log \nu_i &= -\mathbf{X_i}'\delta, \\ p_i &= 1 - GEV(-x_i'\alpha; \xi), \end{aligned}$$

where $i = 1, 2, \ldots, n$.

4.1 Data description

The dataset is shared by Mao et al. (2019), requested from the state department of transportation. This dataset was collected in the State of Washington from 2012 to 2015, over 5238 short road segments. The total number of crashes was 32,298 for a total of 10,894,920 passing vehicles, resulting in the average crash rate being $2.96 \times 10-3$ crashes/passing vehicle. Here we treat exposure as the number of passing vehicles. There is 59.9 percent of zero responses in the dataset. We split the dataset into two segments, 80% used for model training and the other 20% used for testing prediction effect. The number of crashes is the outcome variable, and the covariates are shown in Table 5. We present the frequency, proportion, corresponding crash counts and the number of passing vehicles (exposures) of each level for each covariate of both training portion and testing portion in Table 5.

Total mean of crashes are 6.26 for training data and 5.80 for testing data. Total variance of crashes are 343.90 for training data and 281.07 for testing data. The crash rate is high when the number of intersection is greater than 1 and when there is exit from the highway.

4.2 Model fitting results

Same procedure as described in simulation section was conducted. Besides MAE, DIC and LPML, we add one more criterion, prediction error (PE), which is the mean absolute difference between predicted value on test 20% portion and the true value.

Posterior mean, 95% credit interval of coefficients for Zero Inflated Poisson model were shown in Table 6. β_p is the coefficient for mean parameter and α_p is the coefficient for probability p of zeros. ξ_p is the parameter of GEV link. For each variable, the first level is treated as the reference level. The result of zero inflated COM-Poisson was listed in Table 7. δ_{cmp} is the coefficient for dispersion parameter.

Table 5: Summary statistics for covariates, Training (Testing)

covariates	Frequency	Proportion	Crash counts	exposure	crash rate (10^{-3})	variance
RTE type						
I	1799 (437)	0.429 (0.417)	24079 (5464)	74596405 (18143286)	0.323 (0.301)	699.48 (584.97)
SR	933 (277)	0.223 (0.217)	1259 (323)	5513080 (1432626)	0.228 (0.225)	18.71 (16.63)
US RTE	1458 (384)	0.348 (0.366)	883 (290)	7203335 (2060515)	0.123 (0.141)	3.47 (4.47)
ENTRANCE_EXIT						
0	4131(1032)	0.986 (0.985)	25782 (5958)	86219212 (21408265)	0.299 (0.278)	340.27 (283.60)
1	49 (12)	0.012 (0.011)	149 (96)	951105 (191707)	0.157 (0.501)	40.66 (161.09)
2	10 (4)	0.002 (0.004)	290 (23)	142503 (36455)	2.035 (0.631)	3071.56 (18.25)
Number of intersection						
0	3930 (976)	0.938 (0.931)	25348 (5888)	85938328 (21256677)	0.295 (0.277)	362.25 (298.13)
1	186 (47)	0.044 (0.045)	564 (121)	882526 (229299)	0.639 (0.528)	55.76 (29.68)
2	42 (19)	0.010 (0.018)	189 (65)	239943 (105424)	0.788 (0.617)	79.08 (78.48)
3	12 (2)	0.003 (0.002)	53 (2)	85744 (11835)	0.618 (0.168)	35.90 (2.00)
4	20 (4)	0.005 (0.004)	67 (1)	166279 (33191)	0.403 (0.030)	49.71 (0.25)

Table 6: Zero Inflated Poisson Model Fit Results

	β_p	95% CI of β_p	α_p	95% CI of α_p	ξ_p
Intercept	-7.883	(-7.896, -7.871)	-0.746	(-0.948, -0.585)	-0.04
RTE type					
I					
SR	-0.259	(-0.336, -0.183)	0.526	(0.344, 0.743)	
US RTE	-0.333	(-0.420, -0.248)	0.950	(0.765, 1.157)	
ENTRANCE_EXIT					
0					
1	-0.682	(-0.852, -0.519)	-0.382	(-0.939, 0.077)	
2	1.553	(1.428, 1.674)	-0.082	(-0.758, 0.578)	
Number of intersection					
0					
1	1.354	(1.247, 1.463)	-0.145	(-0.371, 0.074)	
2	1.424	(1.261, 1.582)	-0.408	(-0.853, -0.001)	
3	1.240	(0.953, 1.511)	-0.337	(-0.973, 0.255)	
4	1.034	(0.776, 1.278)	-0.243	(-0.786, 0.252)	
MAE	5.178				
PE	4.873				
DIC	33642.35				
LPML	-4.036				

Table 7: Zero Inflated COM-Poisson Model Fit Results

	β_{cmp}	95% CI of β_{cmp}	α_{cmp}	95% CI of α_{cmp}	δ_{cmp}	95% CI of δ_{cmp}	ξ_{cmp}
Intercept	-9.360	(-9.748, -9.045)	3.692	(3.513, 3.888)	-0.980	(-1.413, -0.692)	0.367
RTE type							
I							
SR	-0.657	(-1.661, 0.106)	-1.881	(-2.238, -1.500)	0.871	(0.562, 1.312)	
US RTE	-3.965	(-6.331, -2.023)	-1.189	(-1.603, -0.787)	1.184	(0.882, 1.614)	
ENTRANCE_EXIT							
0							
1	-2.749	(-5.808, -0.208)	0.633	(0.008, 1.259)	-0.445	(-1.312, 0.121)	
2	0.687	(-2.731, 2.521)	1.524	(-0.313, 3.341)	-0.235	(-1.368, 0.591)	
Number of intersection							
0							
1	2.864	(1.951, 3.885)	0.077	(-0.404, 0.623)	-0.300	(-0.596, -0.042)	
2	1.692	(-1.955, 3.757)	1.101	(0.181, 2.111)	-0.621	(-1.328, -0.108)	
3	0.493	(-3.766, 3.597)	1.295	(-0.045, 2.540)	-0.718	(-1.911, 0.118)	
4	-1.184	(-4.921, 1.977)	1.499	(0.606, 2.388)	-0.562	(-1.512, 0.100)	
MAE	5.622						
PE	5.164						
DIC	15978.35						
LPML	-1.910						

Zero inflated Poisson has slightly smaller MAE and PE than the zero inflated Conway Maxwell Poisson model, but the MAE and PE of both models are very close to each other. In addition, DIC and LPML of zero inflated Conway Maxwell Poisson model has been improved a lot comparing with zero inflated Poisson model. In addition, some coefficients of dispersion parameter of Conway Maxwell Poisson model are significant, meaning the outcome variable is not equi- dispersed, indicating that Poisson model is not a good choice. Overall speaking, zero inflated Conway Maxwell Poisson model is preferred than zero inflated Poisson on this data set.

When we take a look at the coefficient estimates, both models get same sign of the estimated coefficient, except when the number of intersections being 4. The estimated coefficient is negative under zero inflated Conway Maxwell Poisson, which is different from the zero inflated Poisson, but this one is not significant under the 95% credible interval. The coefficient of p, the probability of zeros, of zero inflated Conway Maxwell Poisson also has different sign from the zero inflated Poisson model. The difference might be from the differently estimated skewness parameter ξ of the GEV link. Both estimated ξ are greater than $\log 2 - 1$, meaning the link of the probability p of zeros is positively skewed, which cannot be captured by the usual used logit link.

Conway Maxwell Poisson model can provide more information, which is the dispersion. From the estimated δ, we can get an idea of how dispersed each strata is. For example, the estimated δ of RTE type SR is significantly positive, meaning this strata is more over dispersed than RTE type I.

5 Conclusion

In this paper, we propose a zero inflated Conway Maxwell Poisson model with the flexible GEV link function, with an application on vehicle crashes. We compare the model performance of zero inflated Poisson model using GEV link with logit link and have shown the flexibility of using GEV link in simulation study. In simulation study, we also compare the model performance of zero inflated Conway Maxwell Poisson model with zero inflated Poisson model using GEV link under several scenarios. Based on the simulation study results, we have shown the better performance of zero inflated Conway Maxwell Poisson model. This application demonstrates first time zero inflated COM-Poisson applied on traffic accident data. We consider three predictors, RTE type, whether there is entrance or exit and the number of intersections. From the real data application, we found zero inflated COM-Poisson model performs better than zero inflated Poisson model.

References

Aguero-Valverde, J. (2013), "Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates," *Accident Analysis and Prevention*, 50, 289–297.

Barriga, G. D. and Louzada, F. (2014), "The zero-inflated Conway Maxwell Poisson distribution: Bayesian inference, regression modeling and influence diagnostic," *Statistical Methodology*, 21, 23–34.

Brooks, S. P. and Gelman, A. (1998), "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, 7, 434–455.

Chanialidis, C., Evers, L., Neocleous, T., and Nobile, A. (2018), "Efficient Bayesian inference for COM-Poisson regression models," *Statistics and Computing*, 28, 595–608.

Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999), "A new skewed link model for dichotomous quantal response data," *Journal of the American Statistical Association*, 448, 1172–1186.

Conway, R. W. and Maxwell, W. L. (1962), "A queuing model with state dependent service rates," *Journal of Industrial Engineering*, 12, 132–136.

Cui, Y., Kim, D.-Y., and Zhu, J. (2006), "On the generalized Poisson regression mixture model for mapping quantitative trait loci with count data," *Genetics*, 174, 2159–2172.

Famoy, F., Wulu, J. T., and Singh, K. P. (2004), "On the Generalized Poisson Regression Model with an Application to Accident Data," *Journal of Data Science*, 2, 287–295.

Famoye, F. (1993), "Restricted generalized Poisson regression model," *Communications in Statistics - Theory and Methods*, 22, 1335–1354.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., and Thomsen, Lars Krogsgård, K. (1995), "Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts," *Accident Analysis and Prevention*, 27, 1–20.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Geweke, J. (1992), "Evaluating the accuracy of sampling based approaches to the calculation of posterior moments," in *In Bayesian Statistics*, University Press, vol. 7, pp. 169–193.

Guikema, S. D. and Goffelt, J. P. (2008), "A Flexible Count Data Regression Model for Risk Analysis," *Risk Analysis*, 28, 213–223.

Huang, A. (2017), "Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts," *Statistical Modelling*, 17, 359–380.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.

Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

Lord, D., Guikema, S. D., and Geedipally, S. R. (2008), "Application of the Conway Maxwell Poisson generalized linear model for analyzing motor vehicle crashes," *Accident Analysis and Prevention*, 40, 1123–1134.

Mao, H., Deng, X., Lord, D., Flintsch, G., and Guo, F. (2019), "Adjusting finite sample bias in traffic safety modeling," *Accident Analysis & Prevention*, 131, 112–121.

Nagler, J. (1994), "Scobit: an alternative estimator to logit and probit," *American Journal of Political Science*, 38, 230–255.

Poch, M. and Mannering, F. (1996), "Negative binomial analysis of intersection-accident frequencies," *Journal of Transportation Engineering*, 122, 105–113.

Ridout, M., Hinde, J., and Demetrio, C. G. B. (2001), "A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives," *Biometrics*, 57, 219–223.

Sellers, K. F. and Raim, A. (2016), "A flexible zero-inflated model to address data dispersion," *Computational Statistics and Data Analysis*, 99, 68–80.

Sellers, K. F. and Shmueli, G. (2010), "A Flexible Regression Model for Count Data," *The Annals of Applied Statistics*, 4, 943–961.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005), "A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution," *Journal of the Royal Statistical Society, Applied Statistics*, 54, 127–142.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society*, 64, 583–639.

Wang, X. and Dey, D. K. (2010), "Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption," *The Annals of Applied Statistics*, 4, 2000–2023.

— (2011), "Generalized extreme value regression for ordinal response data," *Environmental and Ecological Statistics*, 18, 619–634.

Received: February 17, 2021

Accepted: April 10, 2021