Applied Stochastic
Models in Business
and Industry

Discussion

(wileyonlinelibrary.com) DOI: 10.1002/asmb.937

Published online in Wiley Online Library

The Conway–Maxwell–Poisson model for analyzing crash data[‡]

1. Introduction

Sellers *et al*. [1] have provided an excellent summary about the flexible and unique properties of the COM-Poisson model both in terms of methodological advancements and applications. This discussion paper further expands on some of these topics, but focuses on the latest research on the use of the COM-Poisson for reducing the negative effects associated with motor vehicle collisions.

The appropriate selection of a model can be quite complex. The selection process is often guided by the study objectives (e.g., prediction versus relationships, etc.), inference goals (e.g., confidence intervals or parameter distribution), and the availability and quality of data. In highway safety (a subfield of transportation engineering), crash datasets (see Section 2) are often characterized by distinctive attributes not commonly found in other areas, such as biology, marketing or environmental engineering. Because of the large costs associated with the data collection procedure, crash datasets often contain a relatively small number of observations [2]. Regression models developed from datasets containing less than 30 observations are not uncommon. Furthermore, because crashes are rare events statistically speaking, many datasets are characterized by low sample mean values, which have been shown to provide biased or unreliable estimates when traditional regression models are used for risk analysis [3]. In short, crash data collected for analysis purposes are considered unique datasets in their own rights.

The application of the COM-Poisson model in highway safety research was initially fostered to examine the model properties for handling crash datasets. So far, the research on the COM-Poisson model has focused on under-dispersion, the observation-specific variance function, and its performance as a function of sample size and sample mean values. Each topic is discussed further below. But first, the characteristics related to crash data are briefly described in the next section.

2. Crash data

Motor vehicle crashes (e.g., collisions between two vehicles, vehicles running-off-the-road and hitting a tree, etc.) are ranked among the 10 leading causes of injuries across the world [4]. In the US, traffic crashes only fall behind cancer and heart disease as the total loss of human life (i.e., human-years) [5]. It is estimated that the economic societal cost caused by motor vehicle crashes was equal to $230.6 billion in 2000 [6]. Given the huge negative impact crashes have on society, the federal and state governments have placed a significant amount of resources on reducing the number and severity of crashes that plague the transportation network.

In highway safety, crash data are considered the best sources for estimating the safety performance of various components of the highway network, such as intersections, highway segments, bridges, and pedestrian crosswalks. These data can provide useful information about potential deficiencies located within the highway network and can consequently be used for allocating resources to implement countermeasures that would reduce the morbidity caused by these events. Crash data can also be used for evaluating the effects of these countermeasures or treatments. Because crash data are discrete, random and non-negative events, various statistical tools are need for conducting different kinds of safety analyses (e.g., identify hazardous sites; determine contributing factors that influence crashes; evaluate

[‡] *Discussion paper associated with 'The COM-Poisson Model for Count Data: A Survey of Methods and Applications' by Sellers, K., Borle, S., and Shmueli, G.*

Copyright © 2012 John Wiley & Sons, Ltd.

countermeasures, etc.). Count data models are still considered the most popular tools for analyzing this type of data [7].

3. Under-dispersion

Although most count datasets are characterized by over-dispersion, it has been found that crash data can sometimes be plagued by under-dispersion. The under-dispersion can be generated by the data generating process (as discussed in [1]) or attributed to the modeling process [8,9]. For the latter, the under-dispersion occurs when the observed count is modeled conditional upon its mean (i.e., marginal model). In some cases, the under-dispersion can be a sign of over-fitting, meaning that the model may contain too many variables. In highway safety, it has been noted that datasets characterized with low sample values sometimes lead to under-dispersion.

To handle under-dispersion, researchers in various fields have proposed alternative models. Sellers *et al.* [1] discussed some of them, including the weighted [10] and the generalized Poisson models [11]. There are also a few other models that have been proposed for handling under-dispersion. The first model is the gamma model, which can handle both over-dispersion and under-dispersion, similar to the COM-Poisson model. Two parameterizations have been proposed. The first parameterization makes use of the continuous gamma function [12], which means that the mean λ cannot be equal to zero (technically speaking, a continuous model or distribution should not be used to analyze count data). This obviously limits its applicability because, in many datasets, it may not be feasible for the mean to be equal to zero (see Ref. [13] about this assumption in highway safety). The second parameterization is based on the gamma waiting-time distribution, which allows for a monotonic increasing or decreasing function [14, 15]. For this parameterization to work, the observations are assumed to be dependent where the observation at time $t-1$ directly influences the observation at time t. For some datasets, this may be possible, but for most datasets, this is not realistic. For instance, a crash that occurred at time t cannot directly influence another one that will occur 6 months after the first event. The second model is known as the Double Poisson model [16]. Similar to the gamma and COM-Poisson models, it can also handle both over-dispersion and under-dispersion, by adding an extra term. This model has rarely been used by other researchers. It is currently being evaluated in the context of crash data analysis by the research team at Texas A&M University and some preliminary results seem to suggest that the distribution has some difficulties handling under-dispersion (compared with the COM-Poisson), but it appears to be more reliable for over-dispersed data.

This is where the COM-Poisson model offers more flexibility over the existing methods for handling under-dispersion. Lord *et al.* [9] have shown that the COM-Poisson performed better than the gamma-waiting time model for data characterized by under-dispersion. Using the data collected at 162 railway–highway crossings (i.e., train–vehicle collisions) located in South Korea between 1998 and 2002 [8], the COM-Poisson provided a better fit and predictive capabilities compared with the gamma-waiting time parameterization. Table I, taken from the original paper, summarizes the comparison analysis.

Table I. Parameter estimates with three different distributions (MLE) [25].

Variables	COM-Poisson	Poisson	Gamma
Constant	–6.657 (1.206)[a]	–5.326 (0.906)[a]	–3.438 (1.008)[a]
Ln(Traffic Flow)	0.648 (0.139)	0.388 (0.076)	0.230 (0.076)
Average daily railway traffic	—[b]	—	0.004 (0.0024)
Presence of commercial area	1.474 (0.513)	1.109 (0.367)	0.651 (0.287)
Train detector distance	0.0021 (0.0007)	0.0019 (0.0006)	0.001 (0.0004)
Time duration between the activation of warning signals and gates	—	—	0.004 (0.002)
Presence of track circuit controller	–1.305 (0.431)	–0.826 (0.335)	—
Presence of guide	–0.998 (0.512)	—	—
Presence of speed hump	–1.495 (0.531)	–1.033 (0.421)	–1.58 (0.859)
Shape Parameter (ν_0)	2.349 (0.634)	—	2.062 (0.758)
AIC	210.70	196.55	211.38
MPB	–0.007	0.004	0.179
MAD	0.348	0.359	0.459
MSPE	0.236	0.252	0.308

[a] Standard error.

[b] The variable was not statistically significant at the 15% level.

AIC, Akaike Information Criterion; MPB, mean pediction bias; MAD, mean absolute deviance; MSPE, mean squared predicted error.

4. Observation-specific variance function

As discussed by Sellers *et al.* [1], the COM-Poisson model is very flexible for modeling the dispersion. The model can be estimated using a constant dispersion or can allow the dispersion to vary for different observations (observation-specific). The latter parameterization in the context of a regression model was initially proposed by Guikema and Coffelt [17, 18]. The parameterization for the mean and the observation-specific shape parameter is described in Equations (1) and (2):

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \tag{1}$$

$$\ln(\nu) = \gamma_0 + \sum_{j=1}^{q} \gamma_j z_j \quad , \tag{2}$$

where x_i and z_j are covariates, with p covariates used in the centering link function and q covariates used in the shape link function (the sets of parameters used in the two link functions do not necessarily have to be identical); μ is a centering parameter that is approximately the mean of the observations in many cases (recall that $\mu = \lambda^{1/\nu}$, as discussed in [14]); and, ν is defined as the shape parameter of the COM-Poisson distribution and is used for estimating the variance.

Recent research in highway safety has shown that the dispersion parameter of the NB model can potentially be dependent upon the covariates of the model and could vary from one observation to another [19–21]. This characteristic has been shown to be observed when the mean function is mis-specified [22], although this may not be true in all cases [23]. When this occurs, the dispersion parameter of the model is considered to be structured (because the coefficients linking the dispersion/shape parameter to the covariates are statistically significant) meaning that it is dependent upon the characteristics of the data [21]. For instance, the dispersion parameter, hence the variance observed in the crash count, could vary geographically, but is not captured in the modeling process. The curious reader is referred to Ref. [21] for additional information about attributes associated with the observation-specific variance function of NB models.

Given this important characteristic, it became essential to examine whether the variance that varied across observations was the same between the NB and the COM-Poisson models. That is, are the observations with a large or small variance the same between both models? Geedipally and Lord [24] examined this characteristic using two crash datasets and found that the variance function estimated for both models were almost the same. For the NB, the variance was calculated using the traditional function $Var\,[Y_k] = \lambda_k + \alpha_k \lambda_k^2$, where μ_k is the estimated mean value and α_k is the estimated dispersion parameter specific for observation k. For the COM-Poisson, the variance can be estimated using the following relationship $Var\,[Y_k] \approx \mu_k / \nu_k$. Because this variance function is based on an approximation, the variance was estimated using the methodology proposed by Francis *et al.* [25]. For a given μ_k and ν_k, 100,000 observations were simulated and the sample mean and variance were calculated from the simulated data.

Figure 1, which is based on crash data collected at 868 signalized intersections in Toronto, Ont. in 1995, shows the comparison analysis for the variance estimated by the NB and COM-Poisson models. The models linked the number of crashes

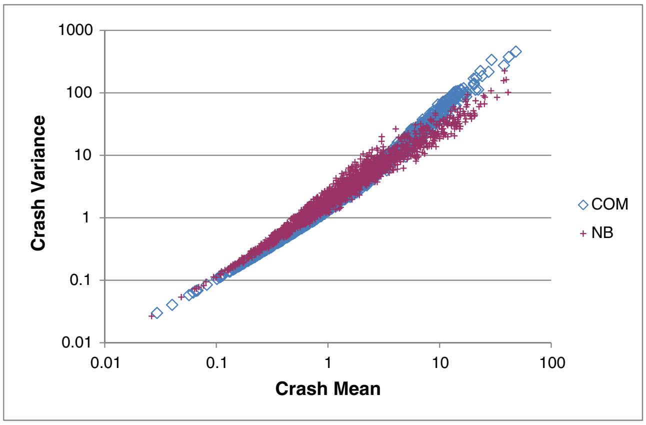

Figure 1. Crash variance versus crash mean for the Texas data [38] (Note: x-axis and y-axis are formatted under a logarithmic scale).

Copyright © 2012 John Wiley & Sons, Ltd.

to the entering traffic flows, defined as major and minor legs, at the intersections: : μ_k (or λ_k) $= \beta_0 F_{k\_\mathrm{major}}^{\beta_1} F_{k\_\mathrm{minor}}^{\beta_2}$; $\alpha_k = \delta_0 F_{k\_\mathrm{major}}^{\delta_1} F_{k\_\mathrm{minor}}^{\delta_2}$; and, $\nu_k = \gamma_0 F_{k\_\mathrm{major}}^{\gamma_1} F_{k\_\mathrm{minor}}^{\gamma_2}$. The study results showed that the COM-Poisson model was able to capture the same variance as for the NB model.

5. Model performance

For a model to be useful, it needs to be reliably and robustly estimated. In other words, the error and bias associated with the model's coefficients should be minimized. It is well-known that the coefficients of models can be strongly and

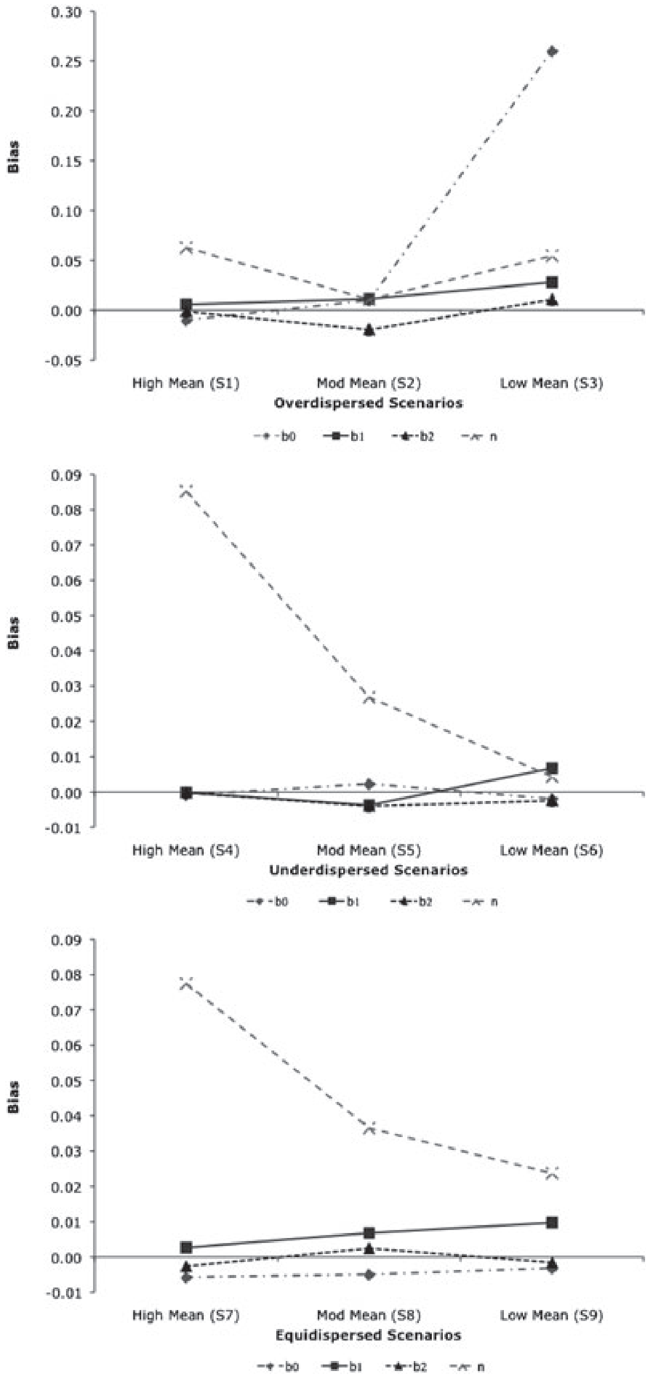

Figure 2. Prediction bias for parameter estimates under each scenario (absolute value). $\boldsymbol{\beta}_0$ diamonds, $\boldsymbol{\beta}_1$ squares, $\boldsymbol{\beta}_2$ triangles, $\boldsymbol{\nu}$ indicated by "x" on the line [25].

Copyright © 2012 John Wiley & Sons, Ltd.

negatively influenced by small sample sizes and low sample means, characteristics often observed with crash data. Several researchers have noted that traditional models, such as the NB model, provide unreliable and biased estimates when they are estimated with such datasets [3, 26–29]. Although no model is immune to these problems, some models may perform better than others under extreme circumstances. For instance, it has been reported that the Poisson-lognormal model is less affected than the NB model for the problems described above [30].

The issue of model performance for the COM-Poisson model has recently been investigated by Francis *et al*. [25]. The authors analyzed the estimation accuracy of the model proposed by Guikema and Coffelt [17, 18] for datasets characterized by over-dispersion, equidispersion and under-dispersion with different mean values based on the maximum likelihood estimator. The analysis was based on 900 simulated datasets representing nine different mean-variance relationships. The functional form of the model was $\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ where x_1, x_2 follows a uniform distribution on [0, 1]. The results showed that the COM-Poisson model was very accurate for all datasets with high and medium sample means irrespective of the type of and level dispersion observed in the data (see Figure 2). However, the model suffered from important limitations for moderate-mean and low-mean over-dispersed data, similar to the NB model, as the sample mean decreased. Interestingly, the model was much less affected when the data are under-dispersed, further suggesting that it is a viable alternative to the other models that have been proposed for modeling under-dispersion.

6. Concluding thoughts

Sellers *et al*. [1] have discussed how the COM-Poisson model has been successfully applied in a variety of fields and its flexibility for analyzing different types of data whether they are over-dispersed, under-dispersed or censored. Bearing in mind that the distribution was first used in a regression setting in the 2003–2005 time frame (see [1]), 'reintroduced' in 2005 [31, 32], and the GLM more fully developed a year later [17], what is very surprising is how quickly the distribution and the model have been used by analysts all around the world and further expanded by researchers over the last four or five years. A quick look at Scopus shows that more than 22 papers have been published in peer-reviewed journals and international conferences on the COM-Poisson (this number does not include forthcoming papers discussed above). All these papers have a combined citation record equal to 114 (at the time this discussion paper was written). This is no small feat, especially considering that a few existing models introduced several years ago (e.g., Double Poisson, etc.) have barely been used by scientists and researchers.

In highway safety, the COM-Poisson model has been applied for different conditions, such as for over-dispersed and under-dispersed data and when the variance is dependent upon the structure of the data. So far, the work on the model has been spearheaded by a small group of researchers, but others have recently shown an interest in using and further developing the COM-Poisson model for safety analyses. The work so far has shown that the COM-Poisson model offers a viable alternative to traditional models for exploring the unique characteristics associated with motor vehicle crashes. However, as discussed by Sellers *et al*. [1], there remain several opportunities for sustained theoretical and applied research related to the COM-Poisson distribution and the regression in highway safety and other fields.

DOMINIQUE LORD
Zachry Department of Civil Engineering, Texas A&M University,
3136 TAMU, College Station, TX 77843-3136, USA
E-mail: *d-lord@tamu.edu*

SETH D. GUIKEMA
Department of Geography and Environmental Engineering, Johns Hopkins University
313 Ames Hall, 3400 N. Charles St., Baltimore, MD 21218, USA

References

1. Sellers K, Borle S, Shmueli G. The COM-Poisson Model for Count Data: A Survey of Methods and Application. *Applied Stochastic Models in Business and Industry* 2012.
2. Lord D, Bonneson JA. Calibration of Predictive Models for Estimating the Safety of Ramp Design Configurations. *Transportation Research Record* 1908; **2005**:88–95.
3. Lord D. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. *Accident Analysis and Prevention* 2006; **38**(4):751–766.
4. WHO. World Report on Road Traffic Injury Prevention. Eds. Peden et al. World Health Organization, Geneva, 2004. (accessed on Sept. 28,2011:http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/).
5. NHTSA. Traffic Safety Facts, 2004 Data. National Highway Traffic Safety Administration, Washington, D.C., 2005. (accessed on Sept. 8, 2011: http://www-nrd.nhtsa.dot.gov/Pubs/809911.PDF).

Copyright © 2012 John Wiley & Sons, Ltd.

6. Blincoe L, Seay A, Zaloshnja E, Miller T, Romano E, Lutchter S, Spicer R. The Economic Impact of Motor Vehicle Crashes. *Report No. DOT HS 809 446*, National Highway Traffic safety Administration, Washington, D.C., 2000. (accessed on Sept. 8, 2011: http://www.cita-vehicleinspection.org/Portals/cita/autofore_study/LinkedDocuments/literature/NHTSA%20the%20economic%20impact%20of%20motor%20vehicle%20crashes%202000%20USA%202002.pdf).
7. Lord D, Mannering F. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research - Part A* 2010; **44**(5):291–305.
8. Oh J, Washington SP, Nam D. Accident prediction model for railway-highway interfaces. *Accident Analysis and Prevention* 2006; **38**(2):346–356.
9. Lord D, Geedipally SR, Guikema S. Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion. *Risk Analysis* 2010; **30**(8):1268–1276.
10. del Castillo J, Pérez-Casany M. Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference* 2005; **134**:486–500.
11. Consul PC. *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker: New York, 1989.
12. Daniels S, Brijs T, Nuyts E, Wets G. Explaining variation in safety performance of roundabouts. *Accident Analysis and Prevention* 2010; **42**(6):1966–1973.
13. Lord D, Washington SP, Ivan JN. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 2005; **37**(1):35–46.
14. Cameron AC, Trivedi PK. *Regression Analysis of Count Data*. Cambridge University Press: Cambridge, U.K., 1998.
15. Winkelmann R. Duration Dependence and Dispersion in Count Data Models. *Journal of Business and Economic Statistics* 1995; **13**:467–474.
16. Efron B. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* 1986; **81**:709–721.
17. Guikema SD, Coffelt JP. An application of the Conway-Maxwell-Poisson generalized linear model. *Society for Risk Analysis Annual Meeting*, Baltimore, MD, December 2006.
18. Guikema SD, Coffelt JP. A flexible count data regression model for risk analysis. *Risk Analysis* 2008; **28**(1):213–223.
19. Heydecker BG, Wu J. Identification of Sites for Road Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. *Advances in Engineering Software* 2001; **32**:859–869.
20. Hauer E. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis and Prevention* 2001; **33**(6):799–808.
21. Miaou S-P, Lord D. Modeling Traffic-Flow Relationships at Signalized Intersections: Dispersion Parameter, Functional Form and Bayes vs Empirical Bayes. *Transportation Research Record* 1840; **2003**:31–40.
22. Mitra S, Washington SP. On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis and Prevention* 2007; **39**(3):459–468.
23. Geedipally SR, Lord D, Park B-J. Analyzing Different Parameterizations of the Varying Dispersion Parameter as a Function of Segment Length. *Transportation Research Record 2103* 2009:108–118.
24. Geedipally SR, Lord D. Examining the Crash Variances Estimated by the Poisson-Gamma and Conway-Maxwell-Poisson Models. *Transportation Research Record* **2011**, in press.
25. Francis RA, Geedipally SR, Guikema SD, Dhavala SS, Lord D, Larocca S. Characterizing the Performance of the Conway-Maxwell-Poisson Generalized Linear Model. *Risk Analysis* 2012. DOI: 10.1111/j.1539-6924.2011.01659.x.
26. Clark SJ, Perry JN. Estimation of the Negative Binomial Parameter κ by Maximum Quasi-Likelihood. *Biometrics* 1989; **45**:309–316.
27. Piegorsch WW. Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter. *Biometrics* 1990; **46**:863–867.
28. Dean CB. Modified Pseudo-Likelihood Estimator of the Overdispersion Parameter in Poisson Mixture Models. *Journal of Applied Statistics* 1994; **21**(6):523–532.
29. Wood GR. Generalised linear accident models and goodness of fit testing. *Accident Analysis and Prevention* 2002; **34**(4):417–427.
30. Lord D, Miranda-Moreno LF. Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective. *Safety Science* 2008; **46**(5):751–770.
31. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Part C* 2005; **54**:127–142.
32. Kadane JB, Shmueli G, Minka TP, Borle S, Boatwright P. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Analysis* 2006; **1**:363–374.

Copyright © 2012 John Wiley & Sons, Ltd.