ELSEVIER

Accident Analysis and Prevention 39 (2007) 53–57

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Further notes on the application of zero-inflated models in highway safety

Dominique Lord [a,*], Simon Washington [b,1], John N. Ivan [c,2]

[a] Zachry Department of Civil Engineering, TAMU3136, Texas A&M University, College Station, TX 77843-3136, USA
[b] Department of Civil & Environmental Engineering, Ira A. Fulton School of Engineering, Arizona State University, USA
[c] Department of Civil & Environmental Engineering, University of Connecticut, 261 Glenbrook Road, Unit 2037, Storrs, CT 06269-2037, USA

Received 10 March 2006; received in revised form 19 May 2006; accepted 13 June 2006

Abstract

The intent of this note is to succinctly articulate additional points that were not provided in the original paper (Lord et al., 2005) and to help clarify a collective reluctance to adopt zero-inflated (ZI) models for modeling highway safety data. A dialogue on this important issue, just one of many important safety modeling issues, is healthy discourse on the path towards improved safety modeling. This note first provides a summary of prior findings and conclusions of the original paper. It then presents two critical and relevant issues: the maximizing statistical fit fallacy and logic problems with the ZI model in highway safety modeling. Finally, we provide brief conclusions.

Keywords: Zero-inflated models; Statistical models; Poisson; Negative binomial; Statistical methods

1. Introduction

Over the last 10 years, the application of zero-inflated (ZI) regression models by transportation safety modelers has gained popularity. Zero-inflated Poisson (ZIP) and negative binomial (ZINB) models have been principally applied when crash data are characterized by a preponderance of zeros. In other words, the data contain more zeros than are expected under a Poisson or negative binomial (NB) (aka Poisson-gamma) distribution given the sample mean. The underlying assumption of ZI models is that entities (e.g., intersections, segments, crosswalks, etc.) exist in two states: a true-zero or inherently safe state (although in recent years some have started defining it as "virtually safe state" to avoid having to defend the notion that sites can be perfectly safe) and a non-zero state (which may happen to record zero accidents in an observation period) that follows the Poisson or NB distribution. Transportation safety analysts have typically justified the use of ZI models because of the improved statistical fit compared to traditional Poisson and NB models (Shankar et al., 1997; Lee and Mannering, 2002; Kumara and Chin, 2003; Shankar et al., 2003). The Vuong statistic (Vuong, 1989) is often a crucial measure of whether the ZIP or ZINB offers a better statistical fit for the modeled data. In fact, very few researchers have solely justified their use based on a dual-state data generating process. If the initial assumption is one of a dual-state process, there is no need to compare the statistical fit of a dual-state model with a single-state model, such as an univariate Poisson or NB regression model (e.g., recent studies where the comparison of distributions was a key analysis component of the study include: Yau et al., 2003; Joe and Zhu, 2005, among others).

Prior to and since the publication of a paper criticizing the use of ZI models in highway safety (Lord et al., 2005), the authors have been contacted by numerous individuals to discuss points described in the paper, in many cases resulting in lively and energetic discussions. The intent of this brief note is to succinctly articulate some additional points (not provided in the original paper) and help to clarify our collective reluctance to adopt ZI models for modeling highway safety data. It is believed that a dialogue on this issue is healthy discourse on the path towards improved safety modeling. This note first provides a *summary of prior findings and conclusions* of the original paper (Lord et al., 2005). It then presents two critical and relevant issues: *the maximizing statistical fit fallacy* and *Logic problems with the ZI model in highway safety modeling*. Finally, we provide *brief conclusions*.

* Corresponding author. Tel.: +1 979 458 3949; fax: +1 979 845 6481.
E-mail addresses: d-lord@tamu.edu (D. Lord), simon.washington@asu.edu (S. Washington), johnivan@engr.uconn.edu (J.N. Ivan).
[1] Tel.: +1 480 965 2220.
[2] Tel.: +1 860 486 0352.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.06.004

2. Summary of prior findings and conclusions (Lord et al., 2005)

Our original paper criticized the use of ZI models in highway safety modeling. Using theoretical principles, simulation experiments, and empirical data cited in the literature, we reported that crash data characterized by a preponderance of zeros is not caused by a dual-state process (i.e., the mixture of truly safe with unsafe sites), but rather is brought about by one or more of the four following conditions:

(1) analysis sites are characterized by a combination of low exposure, high heterogeneity, and sites categorized as high risk;
(2) analyses are conducted with short time or small spatial scales;
(3) the sample data contain a relatively high percentage of missing or mis-reported crashes; and
(4) critical variables are omitted from crash prediction models.

In short, we argued that zero-inflated models do not provide a defensible approach for modeling motor vehicle crashes, even when crash data appear to include a preponderance of zeros. Some of these issues identified above have in fact been corroborated by researchers in other fields who have criticized the application of ZI models when the studied data do not warrant the application of such models (e.g., see Warton, 2005). As reported by the Association of Professional Engineers of Ontario, PEO (1997), and Hauer (1999), inherently safe highways do not exist. Indeed, they maintain that it is inappropriate to argue that a road is safe insofar as crashes are bound to occur. As a result, they suggest that one should never claim a road to be safe, but one should rather refer a highway as either being more or less safe than another highway; in other words, refer to the safety performance of a highway in relative terms.

To circumvent the problems associated with using crash data characterized by a large number of zeros, we offered analytical alternatives that we argued are more theoretically defensible for modeling such datasets. These solutions included changing the spatial or time scale of analysis, including unobserved heterogeneity terms in NB and Poisson models, improving the set of explanatory variables, and applying small area statistical methods (since almost all zero-inflated models described the safety literature have been applied in rural areas where exposure is very low).

It should be noted that this note considers work described in recently published documents on this subject (Shankar et al., 2004; Kumara and Chin, 2005, 2006); interestingly, some of which seem to imply that we approve the application of ZI models for highway safety applications (e.g., Hermans et al., 2006). These recent publications suggest that healthy dialogue on the subject is warranted, and given the abundance of highway safety modeling applications continuing to be reported in the literature, the discussion will likely be ongoing.

3. The maximizing statistical fit fallacy

The prime justification for the use of ZI models has rested on improved statistical fit compared with traditional Poisson and Poisson-gamma models. Despite the fact that the application of such model forms did not seem logical (in the words of some authors), one might argue that this type of model provides improved fit for modeling crash data characterized by a preponderance of zeros (especially when no other alternatives readily exist). The relative importance of statistical fit needs to be put into perspective.

First and foremost, statistical modeling in general is not solely about maximizing statistical fit (e.g., Miaou and Lord, 2003). In fact, this is an incomplete view of the objective of good modeling. [*Note*: As noted by Myers, 1990, "statistics are rarely a substitute for sound scientific knowledge and reasoning. Statistical procedures are vehicles that lead us to conclusions; but scientific logic paves the road along way. However, a scientist must remember that to arrive at an adequate prediction equation, balance must be achieved that takes into account what the data can support . . . For these reasons, a proper marriage must exist between experienced statisticians and learned expert in the discipline involved," p. 165.] The goal of statistical modeling in general is to achieve model parsimony: to maximize fit while simultaneously minimizing complexity. This is the reason for penalized fit measures, such as adjusted R-square. Thus, likelihood ratios can only be improved with the addition of additional model parameters, similar to R-square. Penalized measures such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) used in Bayes' models are examples of measures that penalize for increasing model complexity.

Consequently, Vuong's statistic has been used repeatedly to justify the selection of ZI models. Voung's statistic, however, as reported in much of literature (Washington et al., 2003; Greene, 2000) as a goodness of fit measure does not apply a penalty for additional model parameters. Vuong even identifies this shortcoming and suggests the use of a corrected test statistic using something like AIC or BIC, so that penalties are assigned to models with larger numbers of parameters (Vuong, 1989).

A ZI model that applies a logistic splitting model for zeroes as a function of three covariates, for example, adds four additional model parameters, not to mention additional technical complexity. Thus, the additional complexity added by the ZI model needs to be taken into account when justifying the model from a purely statistical fit perspective (we discuss concerns beyond statistical fit in a moment).

Secondly, maximizing statistical fit is largely a trivial problem. Statistical models in general can be specified quite easily to produce extremely good or even perfect fit to observed data. Adding second, third, and all higher-order interactions for example will produce perfect fit in a linear regression model. One could easily estimate a regression tree using crash data and obtain perfect fit by 'growing' the regression tree so that only a single crash remained at each terminal node. Thus, obtaining good fit is a trivial exercise and should generally not be the prime argument for preferring one model to another. Models should be

chosen based on model parsimony and agreement with theoretical expectations.

This is an especially important consideration in cases when the estimated model is used for predicting crash counts at locations other than those in the estimation data set. It is quite easy to overfit the model to the estimation data, resulting in a model that performs quite poorly 'outside' the estimation data.

4. Logic problems with the ZI model in highway safety modeling

As stated previously, ZI models assume that the phenomenon under study follows a dual-state process: a true-zero and a non-zero state. To illustrate, consider responses to the survey question: "How many times per month do you visit the gambling casino on an average?" There will be a mixture of responses—those who never gamble (true-zeroes) and those who gamble (non-zeroes). For those who do gamble it is possible to observe a zero, and so observed zeroes arise from two underlying states. In highway safety, this translates to an entity (assuming the output variable is the number of crashes per unit of time) that exists either in an inherently safe or non-safe state. In other words, a percentage of entities, which experienced zero collisions during the study period, could be classified as either safe or unsafe. Typically, the probability that an entity belongs to the zero state is estimated using a binary or logistic regression model that may be a function of covariates. All entities that recorded at least one collision during the same observation period will automatically be classified as unsafe (non-zero state). Assuming for the moment that crashes are derived from a dual-state process, several questions surrounding the underlying logic arise:

- What are the boundary conditions delimiting the two states? In all the documents that reported the use of such models, including the ones pioneered by researchers in econometrics (Lambert, 1992; Zorn, 1996; Li et al., 1999), no discussion describing the boundary conditions delimiting the two states have been proposed.
- If the site-specific traits that classify the two states are unobserved (i.e., not present in the observed data), what might they be?
- If one could accurately define a dual-state data generating process, including the appropriate boundary conditions, why analyze these two states together using a single model and estimate a probability that the traits exist in either of the two states via a binary model rather than analyzing the states independently?

In highway safety, models developed using a dual-state approach have generally included, with the exception of traffic flow, the physical characteristics of the road to describe the two states (e.g., see the references listed in the first paragraph). In many instances, the same variables are used to describe their relationship with crashes for the two states simultaneously. For safety applications, the question becomes: what is the boundary delimiting the two states when the physical characteristics of the road are the only attributes describing the state of the entity? Taking an extreme example, assume there are two adjacent highway sections, both recording zero collisions during the observation period and one of which is defined as inherently safe (as identified in the dual-state model). What features or attributes make the first section perfectly safe while the second one is unsafe? When the observed physical characteristics of the road do not change significantly between the two sections (e.g., lane width, shoulder width, pavement type, etc.), the features or attributes must either be unobserved or non-existent. [*Note*: Since both sections are included in the model development, it is assumed that these sections are somewhat homogeneous in their attributes; see Hauer, 2001, and Resende and Benekohal, 1997, for a discussion on the selection of road segments in statistical modelling.] If the features distinguishing these two sections are unobserved then a significant problem arises because we do not know how to identify 'safe' and 'unsafe' sites. If identifying features do not exist then the dual-state process is an artificial construct, applied simply to improve statistical fit. In either case, the application of a dual-state model has left the analyst in a hopeless situation, with little to explain the qualitative difference between inherently safe and inherently unsafe sites.

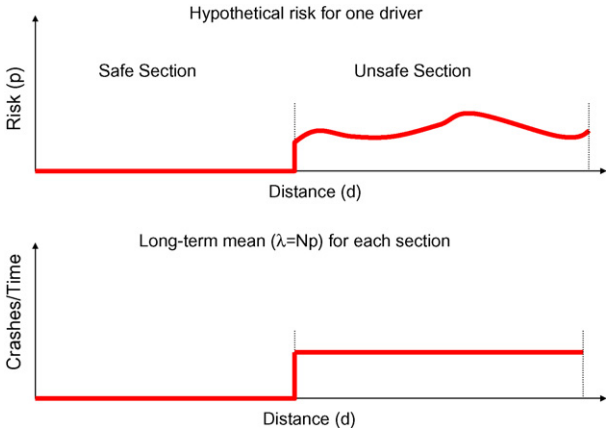

Fig. 1. Crash risk and long-term mean of two hypothetical highway adjacent sections.

A second issue is closely related to the first one. Again, taking the example described previously, given that most of the drivers who traveled through the first section will also go through the second section, it is unreasonable to assume that the crash risk (defined as p, using the notation in Lord et al., 2005) for each driver crossing the first section would equal zero while the risk would suddenly increase for the second section (going from zero to non-zero sites). This hypothetical situation is illustrated graphically in Fig. 1. In this figure, adjacent sections experienced no collision and are assumed to be the same length. The first one is categorized as inherently safe ($Np = \lambda = 0$), while the second one is classified as unsafe ($Np = \lambda > 0$). Since both sections are expected to have somewhat similar physical attributes, it is difficult to explain why one site would have a long-term mean equal to zero while the adjacent site would have mean above zero after all individual risks are cumulated.

A third issue deals with the change in state for the sites under study. One might argue that sites could theoretically change

state. Based on a discussion by Zorn (1996), this would imply that inherently safe sites could on occasion cross the boundary delimitating the two states when the conditions are appropriate and become classified as unsafe. Given the points discussed previously, one is strained to explain the change in state of an entity. For example, if a drunk driver who is traveling on a highway section that has been identified as inherently safe, runs off the road, over-steers, rollovers and is fatally injured, what triggered the change in state, if none of the physical attributes changed prior to the arrival of the drunk driver? Surely the crash outcome itself could not be the 'trigger' changing a site from a safe to unsafe site. Defining the boundary conditions delimiting the two states is critical for explaining why safe sites could potentially become unsafe, if such change exists, and so far such explanations are not forthcoming. Moreover, the overwhelming evidence that accidents are largely random events prevents a site from being immune to crashes by definition (see Klauer et al., 2006, who reported that 80% of crashes are caused by driver inattention).

A fourth issue is related to the use of the output of dual-state models by practitioners, transportation planners, and highway engineers. Previous work on ZI models has claimed that highway sections, intersections, and pedestrian crossings among others can be inherently safe. A practitioner trying to translate such model output may erroneously believe that an entity will be crash (and risk) free. Can one imagine a city engineer at a public meeting claiming that a number of pedestrian crosswalks are inherently safe and risk-free among all crosswalks within his or her jurisdiction? Some may interpret that pedestrians do not need to look for oncoming traffic before initiating a crossing maneuver. Practitioners we have discussed this with agree that such an interpretation is problematic and inconsistent with field-level observations and engineering intuition.

The final issue deals with the importance of selecting the appropriate modeling approach or method for modeling motor vehicle collisions and is somewhat related to the topic discussed in the first section. As discussed previously, the improper use of the modeling method can provide counterintuitive results and have negative consequences on the safety of road users (*Note*: This issue is not limited to ZI models). A good example related to this issue can be found in the application of ZI models at signalized intersections. Using the same dataset that exhibited excess zeros (104 three-legged intersections in Singapore), researchers in England and Singapore used different modeling methods in which the excess zeros were subsequently attributed to inherently safe sites (Kumara and Chin, 2003), was caused by missing values (Kumara and Chin, 2005), and, was finally attributed to heterogeneity in the data (Kumara and Chin, 2006). In this example, one may question which phenomenon appropriately explains the apparent excess of zeros despite the fact that the use of distinct modeling methods lead to different conclusions; as discussed in Lord et al. (2005) and acknowledged in Kumara and Chin (2003, 2005), underreporting was the most likely explanation. Finally, a decision-maker who is only aware of the first study may erroneously make a decision that would affect the safety at signalized intersections located in Singapore.

Table 1
Annual crash counts by severity for a sample of 1-km road segments in Connecticut

Route	1999			2000			2001		
	Fatal	Injury	PDO	Fatal	Injury	PDO	Fatal	Injury	PDO
US 1		9	11		16	16		16	24
CT 14						1			1
CT 176		4	4		2	6		10	14
CT 195		3	10		6	22		6	20
CT 32		4	19		3	23		4	18
CT 44		2	1		1				3
CT 49					3	2			
CT 7		7	11		5	2	1	7	6
CT 85		2	2					4	2

At this point, it is helpful to re-visit the conditions under which apparent "zero inflation" occurs identified previously. Two of these conditions are directly related to issues with the sampling frame and the low expected crash count. In other words, mean crash rates are so low that observations from 1 year to the next at the same location can not only vary by more than 100%, but are guaranteed to result in frequent observations of zero crashes. For example, Table 1 lists crash counts by severity for a sample of 1-km segments on nine Connecticut State Highways. Over the 3 years observed, the property damage only (PDO) crash count varies from 11 to 24 for US 1, and from 2 to 11 for CT 7. Only one of the nine segments had a fatal accident in these 3 years. If we observed CT 49 only in 1999 and 2001, a ZI model would likely call this location a "zero-crash" location, but the 2000 counts show this is clearly not the case. The preponderance of zeros arises when the expected number of crashes in 1 year is very close to zero, but since observed crash counts must be integers, unless a very large number of years is sampled, it is quite likely to observe no crashes in every year sampled. Logically, these locations are not "inherently safe"—they just were not sampled for a long enough time period for any crashes to occur. It is absurd – and potentially dangerous – to call such locations "inherently safe."

5. Conclusions

In summary, this note has presented further discussion points on the application of the ZI models in highway safety. The discussion focused on the rational application of ZI models for modeling motor vehicle crashes. It has been argued on multiple fronts that ZI models, although offering improved statistical fit, should be avoided for modeling motor vehicle crashes on highway entities, especially when other statistical methods, such as small area statistics (Rao, 2003) or extreme value models (Coles, 2001), offer better modeling alternatives for single-state systems. The only possible case we might endorse is when prediction, and only prediction, is the sole research objective (i.e., the model is in essence used as a black box and all input elements that go into the box are unknown to the analyst). Unfortunately, prediction of outcomes is seldom, if ever, the only research objective when estimating crash prediction models, as explanation of findings is of some importance. Moreover, even if

researchers' are concerned only with prediction, consumers of research risk may misinterpret the ZI model, leading to a possible misuse of the output.

This view of ZI models does not mean that this kind of model cannot be used for other applications, as discussed above and in Washington et al. (2003), where the two states and the boundary between the two can be clearly defined. There may exist special or general transportation safety applications where it is appropriate (airports, ports, railroads, etc., but these topics are beyond the scope of the discussion here); however, a logical consistency between the dual-state process and the underlying state of causal processes should exist (see Warton, 2005). Statistical methods and their underlying assumptions need to be applied judiciously in order to achieve model parsimony and to withstand detailed logical scrutiny.

References

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, New York, NY.

Greene, W.H., 2000. Econometric Analysis, 4th ed. Prentice Hall, New Jersey.

Hauer, E., 1999. Safety in Highway Design Standards. Proceedings of the 2nd International Symposium of Highway Geometric Design, San Antonio, TX.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in empirical bayes estimation. Accid. Anal. Prev. 33 (6), 799–808.

Hermans, E., Brijs, T., Stiers, T., Offermans, C., 2006. Impact of Weather Conditions on Road Safety Investigated on Hourly Basis. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.

Joe, H., Zhu, R., 2005. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. Biometrical J. 47 (2), 219–229.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. Report no. DOT HS 810 594, National Highway Transportation Safety Administration, Washington, DC.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 3 (4), 53–57.

Kumara, S.S.P., Chin, H.C., 2005. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transportation Res. Record 1908, 46–50.

Kumara, S.S.P., Chin, H.C., 2006. Disaggregate models to examine signalized intersection crash frequencies. Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.

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

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-road accidents: an empirical analysis. Accid. Anal. Prev. 34 (2), 349–361.

Li, C.-C., Lu, J.-C., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41 (1), 29–38.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Res. Record 1840, 31–40.

Myers, R.H., 1990. Classical and Modern Regression with Applications, 2nd ed. Duxbury Press, Belmont, CA.

PEO, 1997. Highway 407 Safety Review. Professional Engineers of Ontario, Toronto, Canada.

Rao, J., 2003. Small Area Estimation. John Wiley and Sons, Hoboken, NJ.

Resende, P.T.V., Benekohal, R.F., 1997. Effects of Roadway Section Length on Accident Modeling. in: Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities. American Society for Civil Engineers. Chicago, Illinois, pp. 403–409.

Shankar, V.N., Chayanan, S., Sittikariya, Shyu, M.-B., Juvva, N.K., Milton, J.C., 2004. Marginal impacts of design, traffic, weather, and related interactions on roadside crashes. Transportation Res. Record 1897. 156–163.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29 (6), 829–837.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Sci. 41 (7), 627–640.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57, 307–333.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (2), 275–289.

Washington, S.P., Karlaftis, M., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall, Boca Raton.

Yau, K.K.W., Wang, K., Lee, A.H., 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical J. 45 (4), 437–452.

Zorn, C.J.W., 1996. Evaluating Zero-inflated and Hurdle Poisson Specifications. Working Paper. Department of Political Science, Ohio State University, Columbus, OH.