Pergamon

Accid. Anal. and Prev., Vol. 29, No. 6, pp. 829–837, 1997

0001-4575/97 $17.00 + 0.00

PII: S0001-4575(97)00052-3

MODELING ACCIDENT FREQUENCIES AS ZERO-ALTERED PROBABILITY PROCESSES: AN EMPIRICAL INQUIRY

V. SHANKAR, J. MILTON and F. MANNERING*

Department of Civil Engineering, 121 More Hall, Box 352700, University of Washington, Seattle, WA 98195, U.S.A.

(Received 25 November 1996)

Abstract—This paper presents an empirical inquiry into the applicability of zero-altered counting processes to roadway section accident frequencies. The intent of such a counting process is to distinguish sections of roadway that are truly safe (near zero-accident likelihood) from those that are unsafe but happen to have zero accidents observed during the period of observation (e.g. one year). Traditional applications of Poisson and negative binomial accident frequency models do not account for this distinction and thus can produce biased coefficient estimates because of the preponderance of zero-accident observations. Zero-altered probability processes such as the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) distributions are examined and proposed for accident frequencies by roadway functional class and geographic location. The findings show that the ZIP structure models are promising and have great flexibility in uncovering processes affecting accident frequencies on roadway sections observed with zero accidents and those with observed accident occurrences. This flexibility allows highway engineers to better isolate design factors that contribute to accident occurrence and also provides additional insight into variables that determine the relative accident likelihoods of safe versus unsafe roadways. The generic nature of the models and the relatively good power of the Vuong specification test used in the non-nested hypotheses of model specifications offers roadway designers the potential to develop a global family of models for accident frequency prediction that can be embedded in a larger safety management system. © 1997 Elsevier Science Ltd.

Keywords—Accident frequency, Poisson regression, Zero-inflated count models

INTRODUCTION

A significant amount of research has been conducted on the applicability of Poisson and negative binomial (NB) distributions (Miaou, 1994; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1996) to predict accident frequencies (e.g. the number of accidents occurring on a roadway section over some specified time period). The Poisson model is appropriate when the mean and the variance of the accident frequencies are approximately equal and the NB model is appropriate when the data are overdispersed (i.e. the variance of the data is significantly greater than the mean, thus violating a basic property of a Poisson process). While previous accident-frequency research has undoubtedly provided insight into the factors determining accident frequencies, it is important to realize that such traditional applications of Poisson and the NB distributions do not address the possibility that more than one underlying process may be affecting accident frequency likelihoods. In fact, one may view accident frequencies as belonging to two states. One state is when the roadway section from which accident data is being gathered is inherently safe. In theory, no accidents will ever be observed when the roadway section is in this zero-accident state[1]. The second state is an acci-

*Corresponding author. Tel: (206) 543 8935; Fax: (206) 543 1543; e-mail: flm@u.washington.edu

[1]The zero-accident accident state may be an outgrowth of accident severity and accident reporting. Because minor accidents may not be reported, the high number of zero-accident observations may be due to the fact that accidents occurring on a specified section of roadway may not have reached the level of severity that would almost guarantee that they be reported. There is also the possibility of 'near misses' that may indicate a potentially dangerous roadway section even though no accidents have been recorded. It is important to keep these possibilities in mind when interpreting our forthcoming empirical results. That is, the zero-accident state may be truly a zero-accident state or an accident state without severe accidents or just near misses.

dent state where accident frequencies follow some known distribution (e.g. the Poisson or NB distribution). Determining which state the roadway section is in is not straight forward because of the period of observation. For example, suppose accident frequency data are collected for a 1 year period and a given roadway section has no accidents reported. The roadway section could truly be in the zero-accident state or may be in the accident state and just happened to have zero accidents over the 1-year observation period[2]. If such a two-state process is modeled as a single process that assumes that all sections are in the accident state (as is inherently assumed when applying traditional Poisson and NB distributions) the estimated models will be inherently biased because there will be an over-representation of zero-accident observations in the data, many of which do not follow the assumed distribution of accident frequencies. Such estimation can also erroneously suggest overdispersion in the data (i.e. indicating that a NB is appropriate when a Poisson distribution is correct) when the overdispersion is merely an outgrowth of an improperly specified model (i.e. modeling a dual-state system as a single-state system).

Dealing with the possibility of dual-state phenomena can be accomplished using statistical procedures that explicitly recognize the existence of the two states as well as allow for the possibility that these two states may be influenced by different factors. The zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models, both of which seek to account for the preponderance of zero-accident observations often found in accident frequency data, are two approaches that have been applied in a variety of fields, including manufacturing (Lambert, 1992), sociology (Land et al., 1996) and econometrics (Mullahy, 1986), to investigate such dual-state systems. In the area of accident analysis, Miaou (1994) analyzed accident frequencies using the ZIP structure. His research began an exploration, by considering the ZIP structure, of the sources of overdispersion (i.e. from true overdispersion or overdispersion resulting from modeling a dual-state system as a single-state process). However, a complete statistical approach that considers ZIP and ZINB and tests for the correct model structure versus traditional Poisson and NB models has yet to be applied in accident analysis research[3].

This paper explores a number of important issues in the application of zero-altered models to accident frequency analysis. Specifically, we will:

(1) explore the conditions under which the ZIP and ZINB model are more appropriate than simple Poisson and NB models; and
(2) distinguish between spurious NB overdispersion and a true ZIP or ZINB underlying process.

The paper begins by first discussing the analysis methodology. Next, the empirical setting and findings are presented and this is followed by conclusions and recommendations.

METHODOLOGY

Dual-state count-data models, such as the ZIP and ZINB models, explicitly separate the true zero-state process from the parent count-data process and allow for different factors to influence both of these states. In the ZINB model[4] (with an application to accident frequency analysis), this dual-state process is handled by letting Y_i be the number of accidents on roadway section i in some specified time period and letting p_i be the probability that roadway section i will exist in the zero-accident state. Thus $1-p_i$ is the probability that a zero-accident observation actually follows a true NB distribution. Given this

$$Y_i = 0 \text{ with probability } p_i + (1-p_i)\left[\frac{\theta}{\theta+\lambda_i}\right]^{\theta} \tag{1}$$

and

$$Y_i = k \text{ with probability } (1-p_i) \times \left[\frac{\Gamma(\theta+k)u_i^{\theta}(1-u_i)^k}{\Gamma(\theta)k!}\right], \tag{2}$$

where k is the number of accidents (positive numbers starting from one), $\theta = 1/\alpha$ (with α being the dispersion parameter), and $u_i = \theta/(\theta+\lambda_i)$ with λ_i being the mean.

[2]Examination of zero-accident frequencies is important because previous literature (Lambert, 1992) has shown that slight changes in unobserved accident-inducing factors can cause the accident process to move back and forth between the zero-accident state in which accidents are non-existent (or of a low enough severity so as to be unreported) and the accident state where accidents are possible but not inevitable (includes zeros as an outcome). This is an important concern that will be addressed in this paper.

[3]Miaou's model (Miaou, 1994) is a 'zero-truncated' Poisson model. Heilbron (1989) and Mullahy (1986) provide additional examples of this approach. The idea behind this approach is that there is a binary process that determines whether the frequency count is zero or a positive integer. A traditional limitation in modeling this binary process has been that covariates (variables affecting the binary probability) have not been considered. Work by Lambert (1992) and Greene (1994) address this covariate issue by using a logistic specification. We follow the logistic specification approach in this paper and address the high number of zero-accident observations by examining what portion of these zeros is due to heterogeneity (i.e. 'true' overdispersion) and what portion of it is being generated by the dual-state nature of the process.

[4]The ZIP model derivation is similar and is not presented here.

Note that the dispersion parameter, α, relaxes the Poisson assumption that requires the mean to be equal to the variance by letting $\text{Var}[Y_i] = E[Y_i]\{1+\alpha E[Y_i]\}$ In Eqs. (1) and (2), the probability of being in the zero-accident state p_i is formulated as a logistic distribution such that $\text{logit}(p_i) = \mathbf{G}_i\gamma$ and λ_i satisfies $\log(\lambda_i) = \mathbf{H}_i\beta$, where $\mathbf{G}_i$ and $\mathbf{H}_i$ are covariate vectors, and γ and β are coefficient vectors that do vary across i. The covariates that affect the mean λ_i of the accident state may or may not be the same as the covariates that affect the zero-accident state probability (i.e. p_i). Intuition suggests that the covariate vector $\mathbf{G}_i$, which determines the likelihood of a zero-accident state, may differ from the covariate vector $\mathbf{H}_i$, which determines the accident frequency in the accident state. Alternatively, these covariate vectors may be related to each other by a single, real-value shaped parameter τ. In such a case, a natural parameterization is $\text{logit}(p_i) = \tau\mathbf{B}_i\beta$, where $\mathbf{B}_i$ differs from $\mathbf{H}_i$ in that some covariates that were significant in the count model (i.e. in the vector $\mathbf{H}_i$) may be excluded from the model determining the probability of the zero-accident state because they are insignificant. Thus vector $\mathbf{B}_i$ can be equal to or a subset of vector $\mathbf{H}_i$. Circumstances that may limit the vector of coefficients to be the equal across the zero accident and accident states will likely be dependent on the data being used to estimate the model. We will test for this equality in the empirical portion of the paper.

Another concern in model estimation is the possibility of overdispersion in the data (i.e. the variance being greater than the mean) which is a phenomenon often encountered in the study of accident frequencies. If a true dual-state process exists (i.e. suggesting that a zero-altered model is appropriate) and the model is estimated as a single-state count model, overdispersion may be erroneously indicated (i.e. suggesting a NB is appropriate when a ZIP is actually the correct model). To see how this spurious overdispersion arises, note that the NB has

$$\text{Var}[Y_i] = E[Y_i]\{1+\alpha E[Y_i]\} \tag{3}$$

where α is the overdispersion parameter. For the ZIP model,

$$\text{Var}[Y_i] = E[Y_i]\left\{1+\frac{p_i}{1-p_i}E[Y_i]\right\} \tag{4}$$

Thus the term $p_i/(1-p_i)$ could erroneously be interpreted as α. When applying a zero-altered model, the problem then becomes one of distinguishing the underlying NB or Poisson distribution from the zero-accident probability alteration. A statistical test for this has been proposed by Vuong (1989). The Vuong test is a t-statistic-based test with reasonable power in count-data applications (see Greene, 1994). The Vuong statistic is computed as

$$V = \frac{\bar{m}\sqrt{N}}{S_m} \tag{5}$$

where $\bar{m}$ is the mean with $m = \log[f_1(.)/f_2(.)]$ [with $f_1(.)$ being the density function of the ZINB distribution and $f_2(.)$ is the density function of the parent-NB distribution], and S_m and N are the standard deviation and sample size, respectively. A value >1.96 (the 95% confidence level for the t-test) for V favors the ZINB while a value <-1.96 favors the parent-NB (values in between 1.96 and -1.96 mean that the test is indecisive). To carry out the test, both the parent and zero-inflated distribution need to be estimated and tested using a t-statistic. This test can also be applied for the ZIP(τ) and ZIP cases.

EMPIRICAL SETTING

To investigate the relationship of highway geometrics, regulatory control and traffic characteristics on accident frequencies, available data were collected for three major functional classes by geographic location. Data from principal and minor arterial highways in Western Washington of varying geometric and traffic characteristics and data from collector arterials in Eastern Washington were used in this analysis. The Washington State Department of Transportation (WSDOT) supplied these highway data from its Highway Geometrics databases. The Washington State Patrol (WSP) accident data files provided the accident history for each highway section.

The limits of a section were defined by changes to any geometric or roadway variable (e.g. a new section would be identified when the shoulder width changed from 1.83 to 2.44 m). The section-defining information included changes to district number, urban or rural location, state route number, related roadway type, number of lanes, roadway width, shoulder width, presence of curb or retaining wall, divided or undivided highway, speed, average annual daily traffic, truck percentage, peak hour factors and vertical and horizontal curve characteristics.

For shoulder information, in cases where a section of highway contained curbs or walls, an indicator variable was used for identification because no shoulder width data was maintained for these sections. Horizontal and vertical curve information was identified by the geometric data supplied from the WSDOT highway geometric database. Vertical curves were identified by grade and the presence of an angle point or inflection point. Horizontal curves were identified by the length, radius, central angle, horizontal curve type and the direction of curvature. In

addition to the highway geometric data, the computerized state roadway inventory files were used as a source of traffic volume, truck percentage, peak hour factor, geometric and speed data.

For model estimation, our data included a 2-year summary of accident data, from 1 January 1992, to 31 December 1993.[5] Tables 1–3 present summary statistics for the different roadway functional classes.

ESTIMATION ISSUES

Estimation of the parameters γ and β was conducted using maximum likelihood procedures. The likelihood function for the previously-defined splitting [i.e. $\text{logit}(p_i)=\mathbf{G}_i\gamma$ and $\log(\lambda_i)=\mathbf{H}_i\beta$] is based on the following probability density function for the random variable Y_i (accident frequency):

$$p(Y_i)=(1-p_i)\left[\frac{\Gamma(\theta+k)u_i^{\theta}(1-u_i)^k}{\Gamma(\theta)k!}\right]+Z_ip_i \quad (6)$$

where $Z_i=1$ when Y_i is observed to be zero and $Z_i=0$ for all other values of Y_i. The use of the indicator variable Z_i makes maximization of the log-likelihood function easy and uniform across the entire sample. The log-likelihood function is then simply $\Sigma_i \log(p(Y_i))$. The log-likelihood function is estimated using the gradient/line search approach procedure proposed by Greene (1996).

Before proceeding to the estimation results, it is important to note that the derivation of ZIP and ZINB models assume that the events (in our case the annual frequency of accidents on defined roadway sections, $Y=[Y_1,\ldots Y_n]$) are independent. In our case we use accident frequencies in consecutive years thus inducing some correlation among frequencies which would affect the efficiency of coefficient estimates. To investigate the potential impact of this, yearly data from 1993 were omitted and models were estimated

[5]A review of the data revealed that cross street information was not included at intersections. Because this presented a possible specification problem (omitted variable bias), highway sections containing intersections were not included in our sample. Also, roadway sections that had undergone construction during the study period were identified from construction accidents in the databases and excluded from the data base.

Table 1. Summary statistics of key geometric and traffic variables for principal arterials in Western Washington

Variable	Description	Minimum	Maximum	Mean	SD
Frequency	Annual number of accidents on roadway sections	0	84	0.294	1.090
Length	Length of section (in km)	0.016	18.26	0.097	0.223
Speed	Posted speed limit (in km/h)	32	88	78.620	12.830
AADT	Average annual daily traffic per lane	251	26,415	4534.870	4255.480
Radius	Horizontal curve radius (in m)	10.542	15,060.24	755.717	1005.959
Degree of curve	Degree of horizontal curvature [angle, in °, subtended by a 100 m arc, equal to 18,000/(π × Radius)]	0	543.49	8.937	18.255
Tangent length	Total length between horizontal curves (in km)	0	12.928	0.081	0.400
Sharp curve	(1 if >6.56° of curve, 0 otherwise)	0	1	0.352	0.478
Number of lanes	Total number of lanes in section	1	7	2.618	0.990
Flat section	1 if section grade is 0%, 0 otherwise	0	1	0.063	0.242
Straight section	1 if section has infinite radius	0	1	0.498	0.500
Narrow center right section	(1 if <1.52 m, 0 otherwise)	0	1	0.333	0.471

Table 2. Summary statistics of key geometric and traffic variables for minor arterials in Western Washington

Variable	Description	Minimum	Maximum	Mean	SD
Frequency	Annual number of accidents on roadway sections	0	7	0.090	0.346
Length	Length of section (in km)	0.016	15.376	0.090	0.300
Speed	Posted speed limit (in km/h)	40	88	72.170	12.880
AADT	Average annual daily traffic per lane	187	11,016	1691.930	1537.780
Radius	Horizontal curve radius (in m)	10.843	6903.61	456.798	590.371
Degree of curve	Degree of horizontal curvature [angle, in °, subtended by a 100 m arc, equal to 18,000/(π × Radius)]	0	528.39	26.640	45.039
Tangent length	Total length between horizontal curves (in km)	0	53.440	0.080	0.761
Sharp curve	(1 if >6.56° of curve, 0 otherwise)	0	1	0.492	0.500
Number of lanes	Total number of lanes in section	1	5	2.001	0.177
Flat section	1 if section grade is 0%, 0 otherwise	0	1	0.132	0.339
Straight section	1 if section has infinite radius	0	1	0.407	0.491
Narrow center right section	(1 if <1.52 m, 0 otherwise)	0	1	0.746	0.436

Table 3. Summary statistics of key geometric and traffic variables for collector arterials in Eastern Washington

Variable	Description	Minimum	Maximum	Mean	SD
Frequency	Annual number of accidents on roadway sections	0	6	0.61	0.279
Length	Length of section (in km)	0.016	14.368	0.154	0.240
Speed	Posted speed limit (in km/h)	40	88	84.094	10.163
AADT	Average annual daily traffic per lane	146	10,438	982.778	930.558
Radius	Horizontal curve radius (in m)	0.600	15,060.24	851.737	1437.598
Degree of curve	Degree of horizontal curvature [angle, in °, subtended by a 100 m arc, equal to $18{,}000/(\pi \times \text{Radius})$]	0	2864.79	4.136	46.287
Tangent length	Total length between horizontal curves (in km)	0	22.88	0.128	0.614
Sharp curve	(1 if $>6.56°$ of curve, 0 otherwise)	0	1	0.391	0.488
Number of lanes	Total number of lanes in section	1	4	2.061	0.345
Flat section	1 if section grade is 0%, 0 otherwise	0	1	0.400	0.490
Straight section	1 if section has infinite radius	0	1	0.430	0.495
Narrow center right section	(1 if <1.52 m, 0 otherwise)	0	1	0.640	0.480

using 1992 data only. A comparison of coefficient estimates using a likelihood ratio test indicated no significant differences in coefficients (including the 'dispersion' coefficient). Thus it can be concluded that serial correlation is not likely playing a significant role in our data.

FINDINGS

Results of the maximum likelihood estimation of the various models described previously are presented in Tables 4–6[6]. For the principal-arterial data from Western Washington, it was determined that the parent-NB specification best described the underlying process (i.e. a single-state process), while for the minor-arterial data the ZINB specification proved to be more appropriate (i.e. a two-state process). For collector-arterial data from Eastern Washington, the ZIP(τ) specification with constrained coefficients was found to be appropriate [i.e. constraining $\text{logit}(p_i) = \tau \mathbf{B}_i \beta$ instead of allowing the coefficients to be estimated without an implied relation to the parent count distribution][7]. It is evident that with little design variability existing for principal arterials, which are typically designed to full standards, and moderate weather effects prevalent in Western Washington, zero-accident and accident states observed during the survey period seem to remain relatively stable for roadway section lifetimes. Consequently, the likelihood of a zero-inflated process affecting accident frequencies is minimal. On the other hand, for minor arterials, which exhibit more design variability than principal arterials and are typically not constructed to full standards, some zero-inflated process effect is likely to occur[8]. In comparison, collector arterials, which due to the low proportion of high-frequency counts seem less susceptible to residual overdispersion in the non-zero accident state, result in accident frequencies that are best represented by a ZIP(τ) distribution.

The model results for the NB specification for principal arterial accident frequencies are presented in Table 4. An examination of this table indicates that all variables are statistically significant and of plausible sign[9]. The overdispersion parameter α is statistically significant (t-statistic of 37.714) indicating significant overdispersion in the data. As suspected previously, inherent overdispersion in the data is due to the parent NB process and this was validated when the ZINB specification failed to provide a statistically better fit (the Vuong statistic <1.96, which corresponds to the 95% confidence limit of the t-test).

The model results for the ZINB specification for minor arterial accident frequencies are presented in Table 5. The ZINB specification shown in this table was determined to be the appropriate model for describing annual accident frequencies on minor arte-

[6]The choice of variables was achieved through an exhaustive search of the data base in which variables providing significant improvements in the model's log-likelihood function at convergence were chosen. Variable transformations (e.g. use of natural logarithms) were also explored but because there was no compelling empirical and/or theoretical evidence suggesting that they resulted in a superior model, such transformations were not included in the final specification.

[7]The appropriate model is determined statistically, using likelihood ratio tests. Thus there is no ambiguity as to which model should be used. For example, if the ZINB specification, used for minor arterials in Western Washington, had been replaced by the 'single parameter' ZIP(τ) the model, the model would be misspecified because a statistically invalid constraint would have been used.

[8]With their relatively higher proportion of high-frequency accident counts in comparison to collector arterials, minor arterials are also likely to suffer from residual overdispersion suggesting a NB distribution.

[9]The interpretation of the signs of variables must be made in the context of the contemporaneous nature of the variables included in the model. For example, the finding that the coefficient for grade is negative (implying steeper grades are associated with lower accident frequency) may seem counterintuitive but one has to consider other geometric and traffic conditions that are likely to prevail, concurrently, at such roadway sections and the likely mitigation (e.g. signing, etc.) that may be present.

Table 4. Negative binomial estimation of annual accident frequency for principal arterials in Western Washington

Variable	Coefficient	(t-statistic)
Constant	−3.658	−18.163
Annual average daily traffic (AADT) per lane	0.922E−04	24.423
AADT indicator (1 if AADT per lane exceeds 2500 vehicles; 0 otherwise)	0.399	7.899
AADT (continuous) on long sections (>0.40 km) interaction	0.276E−04	1.575
Truck volume as a percentage of AADT	−0.845E−02	−2.743
Grade in percent	−0.200E−01	−3.919
Steep grade indicator (1 if grade >5%, 0 otherwise)	−0.328	−2.363
Flat section indicator (1 if grade ≤1%, 0 otherwise)	−0.961E−01	−3.634
Grade−lane width interaction indicator (1 if flat section width with <3.46 m lanes, 0 otherwise)	0.228	3.562
Speed limit (in km/h)	−0.177E−01	−8.430
High speed indicator (1 if speed limit is 80 km/h; 0 otherwise)	−0.113	−3.124
Length of section (in km)	2.949	64.429
Long-section indicator (1 if section length >0.40 km; 0 otherwise)	−0.556	−4.381
Section length–high speed interaction indicator (1 if section length >0.40 km and speed limit is 80 km/h)	0.329	1.626
Number of lanes	0.258	19.580
Narrow lane indicator (1 if lane width is <3.46 m; 0 otherwise)	0.146	4.757
Narrow shoulder indicator (1 if right shoulder is ≤1.51 m in width; 0 otherwise)	0.147	3.564
Degree of curve [degree of horizontal curvature-angle, in °, subtended by a 100 m arc, equal to 18,000/(π × Radius)]	−0.504E−02	−2.023
Central angle (in °)	0.322E−02	3.699
Side friction factor (interaction between speed and radius)	0.254	2.157
Curve spacing indicator (1 if section has adjacent back-to-back curve; 0 otherwise)	1.250	19.275
Roadside feature indicator (1 if feature is wall; 0 otherwise)	0.231	6.218
Straight section indicator (1 if straight section; 0 otherwise)	0.721	18.086
α (dispersion parameter)	1.541	37.714
Number of observations	38,578	
Restricted log-likelihood (constant only)	−24,446.25	
Log-likelihood at convergence	−21,658.64	

rials (the Vuong statistic of 5.23 indicated a high probability that a two-state process was present). With separate vectors of coefficients affecting the zero-accident and accident states, both effects are found to be statistically significant and of plausible signs. However, it is interesting to note that flat roadway sections are found to positively affect both the zero-accident and possible-accident states (i.e. increases the probability of no accidents and a higher number of accidents). While the coefficient on the zero-accident state is 3.982 compared to 0.680 for the possible-accident state, the marginal effect was found to be statistically insignificant. It is quite likely that this counter-intuitive same-sign effect of flat sections arises because of high-frequency counts or zero counts on such sections. Interestingly from a design standpoint, we can expect to see greater benefits for zero-accident state sections with the flat section variable.

For collector arterial accident frequencies, a ZIP model was first estimated but the likelihood at convergence was not significantly better than the ZIP with constrained zero-accident state coefficients [ZIP(τ)] model, suggesting that the ZIP(τ) is the appropriate form. The model results for the ZIP(τ) are presented in Table 6. The Vuong specification test showed that the dual-state process was justified (the t-distributed Vuong statistic was 22.806). However, it was noticed that when an NB specification was estimated on the same data set with the same set of coefficients, the over-dispersion parameter α was found to be statistically significant (t-statistic of 4.95) which seems to indicate the appropriateness of the NB distribution at first glance. However, when we examined alternate specifications such as the ZINB model, the overdispersion parameter α turned out to be weak with the corresponding Vuong statistic being

Table 5. Zero-inflated negative binomial estimation of annual accident frequency for minor arterials in Western Washington

Variable	Coefficient	(*t*-statistic)
Zero accident probability state as logistic function		
Constant	1.244	6.158
Annual average daily traffic (AADT) per lane	−0.650E−03	−3.213
Truck volume indicator (1 if double trucks percentage is <4%; 0 otherwise)	−0.729	−2.290
Narrow shoulder indicator (1 if right shoulder is <1.51 m in width; 0 otherwise)	−1.702	−4.362
Curve spacing indicator (1 if section has back-to-back adjacent curve; 0 otherwise)	−2.108	−4.774
Flat section (1 if section grade=0%; 0 otherwise)	3.982	3.311
Straight section indicator (1 if straight section; 0 otherwise)	−1.387	−3.002
Non-zero accident probability state as NB function		
Constant	−1.869	−5.989
Annual average daily traffic (AADT) per lane	0.142E−03	6.267
Peak hour effect (continuous) as a percentage of AADT	−0.683E−01	−2.528
Truck volume–grade interaction (AADT per lane as a continuous variable on grades >5%)	−0.238E−03	−2.603
Traffic volume–lane width interaction (AADT per lane as a continuous variable on lanes 3.51 m or less in width)	−0.868E−04	−3.683
Flat section indicator (1 if 0% grade; 0 otherwise)	0.680	5.650
Degree of curve [degree of horizontal curvature-angle, in °, subtended by a 100 m arc, equal to 18,000/(π × Radius)]	−0.314E−01	−2.883
Straight section indicator (1 if straight section; 0 otherwise)	0.399	2.999
Side friction factor (interaction between speed and radius)	1.212	2.258
Straight section–grade interaction (1 if grade >3% on straight section; 0 otherwise)	0.280	2.501
Grade–curve interaction (1 if degree of curve >0.82 and <6.56° on flat section; 0 otherwise)	−0.603	−2.500
α (dispersion parameter)	1.244	6.158
Number of observations	11,591	
Restricted log-likelihood (constant only)	−3862.894	
Log-likelihood at convergence	−3592.810	
Vuong statistic	5.230	

marginally lower than the required 1.96-normally distributed bound. Table 7 shows the decision rule adopted in selecting the model for collector arterial frequencies using the Vuong statistic and the NB overdispersion parameter α as criteria. Initial estimation using the ZINB–NB comparison indicated that the Vuong statistic was marginally <1.96 while the overdispersion parameter (α) t-statistic was 1.70. Such a scenario is illustrated by the top left cell in the matrix shown in Table 7. It is evident then that with the overdispersion induced by the splitting mechanism, the overdispersion spuriously detected by the parent NB is diminished. Consequently an alternate specification such as the ZIP or ZIP(τ) seems more plausible because it is evident that some simultaneous zero-altered process is at work.

CONCLUSIONS AND RECOMMENDATIONS

The intent of this research is to examine alternate count processes that could potentially explain accident frequencies on roadway sections by separating true zero-accident state sections from possible-accident state sections. As evidenced, several variants of the ZIP/ZINB model are plausible and promising depending on the geographic and functional classification of the roadway section. This research has shown that by understanding the causality underlying accident frequencies on zero-accident versus possible-accident sections, roadway engineers have the opportunity to isolate design control factors affecting zero-accident processes and positive accident processes.

In studying the effects of highway geometrics, regulatory control and traffic characteristics on safety using the ZIP/ZINB variants, surprising insights such as the same-sign effects of design variables on the accident and zero accident states were found. Such insights have significant potential in altering traditional design approaches which assume uni-directional impacts of design variables on accident likelihoods.

Another interesting point of note was that the NB distribution (Shankar et al., 1995; Milton and Mannering, 1996) can spuriously indicate overdisper-

Table 6. Zero-inflated negative Poisson (τ) of annual accident frequency for collector arterials in Eastern Washington

Variable	Coefficient	(t-statistic)
Zero accident probability state as logistic function and non-zero accident probability state as Poisson function (vectors of regressors constrained to be the same)		
Constant	−4.525	−8.213
Annual average daily traffic (AADT) per lane	0.255	6.596
Train truck volume (as a percentage of AADT)	0.821E−01	3.530
Truck volume (as a percentage of AADT)	−0.215E−01	−2.643
Low volume indicator (if AADT per lane is <250 vehicles; 0 otherwise)	−0.576	−2.448
Length of section (in km)	0.312	7.901
Number of lanes	0.325	4.540
Curve radius (in m)	−0.924E−04	−6.841
Sharp curve indicator (1 if degree of curve is ≥6.56°; 0 otherwise)	−0.589	−5.688
Short curve indicator (1 if curve length is <0.16 km; 0 otherwise)	−0.237	−2.472
Curve spacing–sharp curve interaction (1 if curve spacing is ≥0.80 km and degree of curve >6.56°; 0 otherwise)	0.978	1.803
Curve spacing indicator (1 if section has back-to-back adjacent curve; 0 otherwise)	2.058	6.287
Grade (in %)	0.288	3.028
Flat section (1 if grade <1%; 0 otherwise)	0.138	2.139
High speed–cross-section interaction (1 if speed limit >72 km/h on a two-lane cross-section; 0 otherwise)	0.416	3.515
Three lane section indicator (1 if cross-section has three lanes; 0 otherwise)	0.359	1.522
τ	−0.382	−3.022
Number of observations	15,629	
Restricted log-likelihood (constant only)	−3602.458	
Log-likelihood at convergence	−3286.572	
Vuong statistic	22.806	

Table 7. Decision rule for model selection under ZINB–NB comparisons using the Vuong statistic and overdispersion parameter criteria

		NB Overdispersion parameter α (t-statistic)	
		<2	>2
Vuong statistic for ZINB–NB comparison	<|1.96|	ZIP(τ) or ZIP variant as alternative to NB	NB
	>1.96	ZIP(τ) or ZIP variant	ZINB(τ) or ZINB variant

sion when the underlying process actually consists of a zero-altered splitting mechanism. The ZIP/ZINB-variant decision rules seem to be reasonably reliable in examining the potential for alternate mixing distributions.

The models shown in this research demonstrate significant flexibility in specification offered by the splitting mechanism and the τ parameter. However, research in this paper has only shown limited variants of the specification. A more thorough examination of accidents on a national level where environmental and geographic conditions vary significantly may reveal other unique insights on when zero-altered processes may be simultaneously at work with parent count processes. Furthermore, the models shown in this paper are limited to non-intersection roadway sections. An examination of intersection accidents could shed light on intersection characteristics that lend some locations to inherent positive accident likelihoods and build on earlier findings provided by Poch and Mannering (1996).

Finally, it must be acknowledged that the potential for ZIP-variant applications is unlimited especially when we consider the issues of railroad crossing safety and inter-modal safety. The advantage such distributions offer from a specification standpoint provides highway engineers the opportunity to investigate the interaction between intermodal mobility and safety.

REFERENCES

Greene, W. (1996) *LIMDEP, Version 7.0, User's Manual.* Econometric Software, Bellport, NY.

Greene, W. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression

models (Working Paper EC-94-10) Stern School of Business, New York University, New York.
Heilbron, D. (1989) Generalized linear models for altered zero probabilities and overdispersion in count data. (Technical Report) Department of Epidemiology and Biostatistics, University of California, San Francisco.
Lambert, D. (1992) Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* **34**, 1–14.
Land, K. C., McCall, P. L. and Nagin, D. S. (1996) A comparison of Poisson, negative binomial and semiparametric mixed Poisson regressive models with empirical applications to criminal careers data. *Sociological Methods and Research* **24**, 387–442.
Miaou, S. P. (1994) The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis and Prevention* **26**, 471–482.
Milton, J. C. and Mannering, F. L. (1996) The relationship between highway geometrics, traffic related elements and motor vehicle accidents. (Final Research Report, WA-RD 403.1) Washington State Department of Transportation, Washington.
Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.
Poch, M. and Mannering, F. L. (1996) Negative binomial analysis of intersection accident frequency. *Journal of Transportation Engineering* **122**, 105–113.
Shankar, V. N., Mannering, F. L. and Barfield, W. (1995) Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention* **27**, 371–389.
Vuong, Q. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–334.