Paper No. 03-3476

Accident Count Model Based on Multiyear Cross-Sectional Roadway Data with Serial Correlation

Gudmundur F. Ulfarsson and Venkataraman N. Shankar

The use of the negative multinomial model to form a predictive model of median crossover accident frequencies with a multiyear panel of cross-sectional roadway data with a roadway section-specific serial correlation across time was explored. The negative multinomial model specification is compared with previous research, which used the same database but which also used negative binomial and random-effects negative binomial count models. If there is no section-specific correlation in the panel, the negative multinomial model becomes equivalent to the negative binomial. The differences in the estimation results between those models show that such a correlation exists in the data. The results show that the negative multinomial significantly outperforms the negative binomial and the random-effects negative binomial in terms of fit, with a statistically significantly higher likelihood at convergence. The signs of the coefficients were similar in all models; when the signs differed, the negative multinomial model results were more intuitive. Overall, the analysis supports the use of the negative multinomial count model to estimate median crossover accident frequency models that are based on panel data.

Median crossover accidents are generally severe and result in a high cost to society. Such accidents also have a greater potential of creating liability, both because of their severity and because of the inherent link with design deficiencies, that is, no or weak median barriers. The Washington State Department of Transportation (WSDOT) is creating a systematic process for determining median barrier requirements on state highways. Methodologies are also being developed to address roadway safety design issues related to divided highways in a cost-effective manner. WSDOT is reexamining current median barrier installation guidelines, which are ad hoc and which do not make use of current multivariate statistical techniques that can account for effects from roadway geometric factors, traffic, and the environment.

Median barrier requirements do not currently benefit from advanced analysis of median crossover accidents since limited analysis techniques have been used (*1–3*). The current version of the AASHTO *Roadside Design Guide* suggests the use of a simple bivariate analysis of average daily traffic (ADT) and median width as a guide in determining median barrier requirements (*4*). The relationship is a simple decision rule chart; if ADT is above a certain value and the median width is below a certain value, the chart shows the recommended type of median barrier. At present, WSDOT uses a related ADT-versus-median width relationship when examining median barrier requirements.

There is much to gain from accurate predictions of median crossover frequencies. Installation of median barriers effectively reduces median crossover rates to near 0, although there is always a small chance of median barrier penetration. However, it is not beneficial to install median barriers everywhere on the road network because the accident frequency tends to increase in the presence of barriers. Median barriers reduce the area that vehicles have to recover from or escape an accident in the roadway, and they cause rebound accidents when vehicles strike the barrier and rebound to strike another vehicle traveling in the same direction. Generally, median barriers reduce the frequencies of injury accidents, particularly severe accidents. However, contrary examples show that the case is not that simple. For example, a before-and-after study by Seamons and Smith found a total increase in all injury (including fatal injury) accidents of roughly 14% when median barriers were installed at freeway locations (*5*). Median crossover accidents tend to be more serious, with a higher probability of fatalities, whereas barriers cause a greater frequency of accidents, but most are of lesser severity. Median barriers also carry a maintenance cost, which is threefold: direct monetary cost, traffic delays, and risk to road crews. A cost–benefit analysis that considers the whole project in addition to the predictive models is therefore an important part of developing a full picture to help plan effective measures. Departments of transportation attempt to strike a balance when scheduling sections for barrier installation, and the decision making benefits from an accurate prediction of median crossover frequency.

This paper reports on an examination of a predictive model of median crossover accident frequencies in road sections without median barriers with multiyear data, that is, panel data, and a statistical model that accounts for a section-specific serial correlation across time. This study follows up on the study by Shankar et al. (*6*) by studying the same model variables with the same data set but uses a negative multinomial (NM) model for the longitudinal count data (*7*). This allows a direct comparison between the NM model presented and the models currently favored for overdispersed accident counts, the negative binomial family of models. This fits researchers with the tools to implement a NM model in their own setting. The comparison with the previously published and accepted negative binomial models supplies practitioners with evidence of the improvements in model fit, and thereby the predictive power, that can be had by the use of NM models for count data panels.

METHODOLOGY

Several main effects and interactions have been found to be important in determining accident causality and frequency, as reported by previous research, which has focused on the relationship between accident frequency and factors such as roadway geometrics and

Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195.

environmental measures (*6, 8–10*). Driver and vehicle factors play important parts in the causality picture. In count models such factors cannot be used except in the aggregate, for example, the average age of drivers, because the count models are based on aggregated counts over a certain time period at a particular location.

To proceed with a study of median crossover accidents, the roadway network is separated into a number of sections and accident counts are accumulated over a year. The geometric factors of the roadway sections are known. Median crossover accidents occur in lower numbers than many other accident types, so a longitudinal accident frequency history for a number of years is necessary to achieve reasonable numbers of observations; in other words, a longitudinal panel is formed.

The main estimation problem when models are estimated with panel data is serial correlation. There are two types directly associated with panel data: serial correlation among observations from the same section across time and serial correlation among observations from the same time period. Of those, the key correlation is the time series correlation between observations from the same section, in particular, a correlation between successive years with the effect fading over time, whereas in a data set with data aggregated over a long time period (e.g., a year) and widely distributed spatially, the correlation of observations from the same time period can be expected to be weak or nonexistent. Also, a potential spatial correlation across observations from sections that are close to each other geographically is likely to be minimal in this data set because of the distance between sections.

The temporal serial correlation violates the independence assumptions on unobserved error terms and if left unaccounted for causes the coefficient estimates to be inefficient and the estimated standard errors to be biased. Despite the use of a multivariate model, there may still be some unobserved section-specific heterogeneity that, if correlated with any of the included variables, can bias the coefficient estimates. The NM model's handling of the section-specific correlation will work to capture such unobserved section-specific effects.

Previous research has identified that use of the negative binomial (NB) distribution is a good approach to the modeling of accident counts because they are typically overdispersed (*6, 8*). Overdispersion occurs when the variance of the counts exceeds the mean, thereby violating the assumption of numeric equality between variance and the mean inherent in the Poisson model for count data. In addition to overdispersion, the model must handle the violation of the independent observations assumption (because of a serial correlation across time) made by both the Poisson and the NB models for the estimation to remain efficient. Count model counterparts to the fixed-effects and random-effects least-squares models exist. Hausman et al. examined fixed-effects NB (FENB) and random-effects NB (RENB) models for panel data (*11*). The FENB model does not allow for section-specific variation, but the RENB model allows for randomly distributed section-specific variation. In the case of accident frequency, it is likely that section-specific effects will be important. Shankar et al. explored the use of NB, FENB, and RENB models for median crossover accident frequencies and found the performance of the RENB model to be superior to that of the NB model but found the FENB model to be inestimable with their particular data set (*6*). They suggest that use of the NM model could be appropriate because it accounts directly for section-specific correlation and allows for overdispersion (*7, 12*).

This research incorporates these insights given by previous research into the NM modeling framework, which explicitly accounts for the serial correlation across time that affects yearly accident counts in roadway sections and the unobserved section-specific heterogeneity. The reader is referred to the work by Guo for a more detailed derivation than that given in this section and for another example of the application of the NM model (*7*).

The derivation of the NM model begins with the Poisson probability density function, which is an expression for the probability of the frequency equaling a particular count:

$$P(Y_{it} = y_{it}) = \frac{e^{-\lambda_{it}}\lambda_{it}^{y_{it}}}{y_{it}!} \qquad \text{for } y_{it} = 0, 1, 2, \ldots \tag{1}$$

where Y_{it} is the observed frequency of accidents in section i at time t, and λ_{it} is the mean of the number of accidents. The Poisson model is estimated by defining

$$\ln \lambda_{it} = x_{it}\beta \tag{2}$$

where x_{it} is a vector of section- and time-specific explanatory variables, and β is a vector of the coefficients to be estimated. To account for the section-specific variation, one proceeds as is done for the NB model. A random error term is added to the expression for the mean:

$$\ln \lambda_{it} = x_{it}\beta + \epsilon_i \tag{3}$$

where ϵ_i is a section-specific (not observation-specific, as in the NB model) random error term, and $\exp(\epsilon_i)$ is assumed to be independently and identically distributed gamma with a mean of 1 and variance (α) equal to $1/\theta$. The assumption of a mean of 1 does not cause a loss of generality if Equation 3 includes an intercept term.

The conditional joint density function of all individual counts for a particular section i, given that the individual counts are distributed by Equation 1 and conditioned on ϵ_i, can now be written as

$$P(Y_{i1} = y_{i1}, \ldots, Y_{it_i} = y_{it_i} | \epsilon_i) = \prod_{t'=1}^{t_i} P(Y_{it'} = y_{it'} | \epsilon_i) \tag{4}$$

where t_i denotes the number of time periods observed for section i. This assumes that the accident counts in different sections are independent. This is not unreasonable, because these sections are generally not next to each other and will therefore share only minimal unobserved effects. The unconditional joint density function for the NM distribution can now be derived by integrating Equation 4 and by using the assumed distribution of $\exp(\epsilon_i)$ to give

$$P(Y_{i1} = y_{i1}, \ldots, Y_{it_i} = y_{it_i}) =$$
$$\frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_{i1}! \ldots y_{it_i}!}\left(\frac{\theta}{\eta_i + \theta}\right)^{\theta}\left(\frac{\eta_{i1}}{\eta_i + \theta}\right)^{y_{i1}} \cdots \left(\frac{\eta_{it_i}}{\eta_i + \theta}\right)^{y_{it_i}} \tag{5}$$

where

Γ = gamma function,
$\eta_{it} = e^{x_{it}} \cdot \beta$,
$\eta_i = \eta_{i1} + \ldots + \eta_{it_i}$, and
$y_i = y_{i1} + \ldots + y_{it_i}$.

Recall that the variance of $\exp(\epsilon_i)$ is α equal to $1/\theta$. The degenerate case, when each section has only one observation (i.e., there is no section-specific correlation), yields the negative binomial distribution.

The expected value, the variance, and covariance for the NM model are, respectively,

$$
\begin{aligned}
&E(Y_{it}) = \eta_{it} \\
&\mathrm{Var}(Y_{it}) = E(Y_{it})[1 + \alpha E(Y_{it})] \qquad \text{and} \\
&\mathrm{Cov}(Y_{it}, Y_{it'}) = \alpha \eta_{it} \eta_{it'} \qquad (6)
\end{aligned}
$$

A likelihood function is written by using Equation 5 to give Equation 7:

$$L(\beta, \theta) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \theta)}{\Gamma(\theta)} \left(\frac{\theta}{\eta_i + \theta} \right)^{\theta} \left(\frac{\eta_{i1}}{\eta_i + \theta} \right)^{y_{i1}} \cdots \left(\frac{\eta_{it_i}}{\eta_i + \theta} \right)^{y_{it_i}} \qquad (7)$$

where n is the total number of sections. The estimable coefficients β and α equal to $1/\theta$ are estimated by maximizing the likelihood function (Equation 7) (*13*).

RESULTS

The panel data in this research consists of 5 years (1990 to 1994 inclusive) of annual median crossover accident counts for 275 roadway sections in Washington State. The panel is balanced, with all sections having a full 5-year history. Table 1 gives a summary of the main geometric and traffic characteristics of the sections. This panel represents all sections (longer than 800 m) without median barriers on divided state highways. The reasons that sections longer than 800 m are selected are that about 95% of shorter sections on divided highways have barriers and that the shorter sections are more affected by access controls and intersections (*6*).

The explanatory variables used in this research are the same as those used by Shankar et al. (*6*) in their estimation of NB and RENB models to facilitate comparison with the NM model in the present paper and to statistically test if the NM model outperforms the NB or RENB model in terms of likelihood. The model structure is multivariate, with the main variables accounting for volume, median widths, shoulder widths, and horizontal curves. ADT is important because the higher the volume, the greater the exposure and therefore the greater the frequency of median crossover accidents. The model includes indicators for low and medium volumes, which must be interpreted relative to higher volumes. Median width and various interactions between median width and section length are included to put sections of different lengths on a balanced footing. The effect of horizontal curves is accounted for by use of the number of curves per kilometer, since section lengths are different. Furthermore, interactions between the number of horizontal curves and shoulder width deviation are used to account for a varying environment with shoulder widths changing in a section with more than two horizontal curves. A high density of horizontal curves per kilometer and the pavement friction factor are also used as indicators for curved sections with smoother pavements. In addition, there are four indicators for main state routes in Washington State.

The model results of Shankar et al. (*6*) for their best-fitting NB and RENB models are shown in Table 2 along with the present results for the NM model. By exploring the particular results in Table 2, it can first be seen that all but two coefficients have the same signs in all models. The coefficient values are similar but not identical; this indicates that the NM model is not equivalent to the NB model, which shows that there is a serial correlation among observations from the same section across time. Without such a serial correlation, the NM and NB models would have the same likelihood function.

By examining the details of the results in Table 2, the ADT variables show that accident frequencies are predicted to be lower in sections with lower volumes, as expected. The median width between 9.14 and 12.19 m variable is an indicator for median widths that are on the transition between narrow medians, which are typically treated with a barrier, and wider medians, which are harder to cross. As expected, the coefficient is positive, which indicates

TABLE 1 Summary Statistics of Key Geometric and Traffic Variables

Variable	Average	Minimum	Maximum	Standard Deviation
Accident frequency (accidents/year)	0.241	0	7	0.645
Length (km)	3.949	0.8	33.312	4.400
Average daily traffic (ADT)	37,355	3,345	172,555	36,875
Maximum shoulder width (meters)	1.367	0	3.048	0.513
Minimum shoulder width (meters)	1.617	0	4.994	0.737
Median width as indicator variable[a]	3.691	1	5	1.434
Access control[b]	1.811	0	2	0.427
Speed limit (km per hour)	95.477	56	104	8.8
Single-unit truck percentage	4.196	1.9	10	1.215
Double-unit truck percentage	7.762	0.55	17.8	4.621
Truck percentage	14.163	3.2	32	6.682
Number of grade breaks in section	3.866	0	28	4.088
Number of curves in section	2.749	0	29	2.861
Total number of lanes	4.604	2	8	1.132

[a] Median width as indicator variable has the following categorical definitions: 1 – width less than or equal to 9.14 meters, 2 – between 9.14 and 12.19 meters, 3 – between 12.19 and 15.24 meters, 4 – between 15.24 and 18.29 meters, and 6 – greater than 18.29 meters.

[b] Access control codes: 0 – full access control; 1 – partial access control; 2 – no access control.

TABLE 2 NM Model Coefficient Estimation Results for Median Crossover Accident Frequency

Variable	NB	RENB	NM
Constant	−1.551 (0.181)†	−0.118 (0.391)	−1.500 (0.251)†
ADT less than 5,000 vehicles per lane daily, indicator	−1.398 (0.186)†	−1.373 (0.190)†	−1.381 (0.312)†
ADT between 5,000 and 10,000 vehicles per lane daily, indicator	−0.233 (0.158)	−0.266 (0.157)‡	−0.298 (0.290)
Median width between 9.14 meters and 12.19 meters, indicator	0.463 (0.206)†	0.368 (0.215)‡	0.432 (0.309)
Number of horizontal curves per kilometer	−0.309 (0.128)†	−0.325 (0.141)†	−0.502 (0.262)‡
Length of section (km) if median width is less than 12.19 meters, 0 otherwise	0.281 (0.047)†	0.278 (0.062)†	0.175 (0.052)†
Length of section (km) if median width is between 12.19 meters and 18.29 meters, 0 otherwise	0.526 (0.065)†	0.502 (0.070)†	0.292 (0.068)†
Length of section (km) if median width is greater than 18.29 meters, 0 otherwise	−0.358 (0.060)†	−0.343 (0.065)†	0.105 (0.026)†
Difference between max. and min. shoulder width is > 1.22 meters and the number of horizontal curves is greater than two per section, indicator	0.542 (0.321)‡	0.489 (0.285)‡	0.486 (0.580)
Roadway friction factor if number of horizontal curves is greater than 0.67 per kilometer, 0 otherwise	0.011 (0.004)†	0.010 (0.005)†	0.009 (0.006)
Washington State Route 2, indicator	−2.093 (1.098)‡	−1.973 (1.371)	0.271 (0.587)
Washington State Route 16, indicator	−1.338 (0.581)†	−1.290 (0.792)	−1.188 (0.746)
Washington State Route 90, indicator	−0.722 (0.199)†	−0.732 (0.195)†	−0.560 (0.341)
Washington State Route 205, indicator	−1.814 (1.055)‡	−1.756 (1.150)	−8.815 (0.533)†
α	0.447 (0.172)†		0.258 (1.074)
a		128.780 (312.380)	
b		34.514 (90.241)	
ln $L(\beta = 0, \alpha = 1)$, naïve model	**	**	−827.556
ln L at NB values			−883.746
ln L at convergence	− 711.931	− 715.801	−613.078

NOTE: Standard errors are given in parentheses. An "indicator" variable is 1 if the condition holds and 0 otherwise.
The NB and RENB model results presented elsewhere (*6*) are presented here for comparison with the NM model results.
† = significance at the 95% level by the two-tailed *t*-test; ‡ = significance at the 90% level by the two-tailed test;
a,b = parameters of the beta distribution used in the RENB model; ** = information not available.

higher median crossover frequencies. The number of horizontal curves per kilometer has a negative effect on accident frequencies, but note that this is accounting for the positive effect on curves with high deviations in shoulder width and the pavement friction factor. The length variables are all positive in the NM model, but the length of section for wide medians (wider than 18.29 m) was negative in the NB and RENB models. This is a more intuitive result, since the longer the section is, the greater the exposure and the higher the frequency. The NM model results in the opposite sign for the indicator for State Route 2 (SR-2); however, the coefficient is not significantly different from 0; also, the NM model results in a much larger negative coefficient for SR-205. The general result is that the NM model corroborates the result that SR-16, I-90, and SR-205 experience lower median crossover frequencies when the other factors are accounted for. Refer to the work of Shankar et al. for additional discussion of the signs of particular variables in light of the NB and RENB models (*6*).

The results of Shankar et al. were that the NB model outperformed the RENB model when spatial and temporal indicator variables (which captured fixed effects) were included in the NB model but that the RENB model outperformed the NB model when those variables were not included (*6*). They used the likelihood ratio test to test for a significant improvement in log likelihood, the same test used below. They also found that the overdispersion parameter α was significantly different from 0, indicating significant overdispersion and the validity of the NB model. This is not the case for the NM model presented here, for which the dispersion factor is not significantly different from 0. This may indicate that some of the overdispersion may be captured by the NM model's treatment of section-specific serial correlation, thereby rendering the overdispersion factor insignificant. Note that without a section-specific correlation, the NM model likelihood is exactly the same as the NB model likelihood, and with the same data and variables, the only difference is the NM model's handling of the temporal correlation across observations from the same section.

To test statistically whether the NM model outperforms the NB model, the unrestricted log likelihood (ln L_U) at convergence for the NM model (which is equal to −613.078) and the restricted log likelihood (ln L_R) were compared, in which the NM model likelihood function is evaluated for the coefficient estimates of the NB model from Shankar et al. (*6*) by a likelihood ratio test (*13*). The value of ln L_R is −883.746. The likelihood ratio test statistic is

$$L_R = -2(\ln L_R - \ln L_U) = 541.336$$

This is a chi-square (χ^2)-distributed statistic, with the number of degrees of freedom equal to the number of restrictions, which in this case was 15, as all the parameters in the model were restricted to the NB model values. If the likelihood ratio statistic is significantly greater than 0, then the NM model has a significantly higher log likelihood and its estimated coefficients outperform the NB model estimated coefficients in terms of fit and predictive power. The test yields a large likelihood ratio statistic of 541.336, which is a significant difference at a greater than 99.9% ($p < .001$) level of significance. The authors conclude that the NM model gives a significantly superior fit when median crossover accident frequencies from a longitudinal panel of accident counts on roadway sections are modeled.

The results presented here largely support the conclusions of Shankar et al. with respect to the effects of variables (*6*). In particular, they show that median crossover accident frequencies are affected by many factors in addition to median width and average daily traffic. The results also show that NB models, even with temporal and spatial effects, and RENB models are outperformed by the negative multinomial model in terms of likelihood. The negative multinomial model is therefore an appropriate way to form a predictive model of multiyear median crossover count frequencies for a cross-section of roadway sections with serial correlation across time.

ACKNOWLEDGMENTS

This work was supported by the Washington State Department of Transportation Research Project on Median Crossover Accidents. The authors extend special thanks to Yanfeng Ouyang for help in programming the negative multinomial likelihood. The authors also wish to thank the anonymous reviewers for constructive comments that helped improve the paper.

REFERENCES

1. Graf, V. D., and N. C. Winegard. *Median Barrier Warrants.* Traffic Department of the State of California, 1968.
2. Ross, H. E., Jr. *Impact Performance and Selection Criterion for the Texas Median Barriers.* Research Report 140-8. Texas Transportation Institute, Texas A&M University, 1974.
3. Bronstad, M. E., L. R. Calcote, and C. E. Kimball. *Concrete Median Barrier Research.* Report FHWA-RD-73-3. FHWA, U.S. Department of Transportation, June 1976.
4. *Roadside Design Guide.* AASHTO, 1996.
5. Seamons, L. L., and R. N. Smith. *Past and Current Median Barrier Practice in California.* Final Report CALTRANS-TE-90-2. California Department of Transportation, 1991.
6. Shankar, V. N., R. B. Albin, J. C. Milton, and F. L. Mannering. Evaluating Median Crossover Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. In *Transportation Research Record 1635,* TRB, National Research Council, Washington, D.C., 1998, pp. 44–48.
7. Guo, G. Negative Multinomial Regression Models for Clustered Event Counts. *Sociological Methodology,* Vol. 26, 1996, pp. 113–132.
8. Shankar, V. N., F. L. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Conditions on Rural Accident Frequencies. *Accident Analysis and Prevention,* Vol. 27, No. 3, 1995, pp. 371–389.
9. Milton, J. C., and F. L. Mannering. *The Relationship Between Highway Geometrics, Traffic Related Elements and Motor Vehicle Accidents.* Final Research Report WA-RD 403.1. Washington State Department of Transportation, 1996.
10. Shankar, V. N., J. C. Milton, and F. L. Mannering. Modeling Statewide Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention,* Vol. 29, No. 6, 1997, pp. 829–837.
11. Hausman, J. A., B. H. Hall, and Z. Griliches. Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica,* Vol. 52, No. 4, 1984, pp. 909–938.
12. Greenwood, M., and G. U. Yule. An Inquiry into the Nature of Frequency Distribution of Multiple Happenings. *Journal of the Royal Statistical Society, A,* Vol. 83, 1920, pp. 255–279.
13. Greene, W. *Econometric Analysis,* 4th ed. Prentice Hall, Upper Saddle River, N.J., 1999.

Publication of this paper sponsored by Committee on Safety Data, Analysis, and Evaluation.