Pergamon

Accid. Anal. and Prev., Vol. 26, No. 4, pp. 471–482, 1994

0001-4575/94 $6.00 + .00

0001-4575(94)E0010-I

THE RELATIONSHIP BETWEEN TRUCK ACCIDENTS AND GEOMETRIC DESIGN OF ROAD SECTIONS: POISSON VERSUS NEGATIVE BINOMIAL REGRESSIONS

SHAW-PIN MIAOU

Center for Transportation Analysis, Energy Division, Oak Ridge National Laboratory P.O. Box 2008,
MS 6366, Building 5500A, Oak Ridge, TN 37831, U.S.A.

(*Accepted* 25 *September* 1993)

Abstract—This paper evaluates the performance of Poisson and negative binomial (NB) regression models in establishing the relationship between truck accidents and geometric design of road sections. Three types of models are considered: Poisson regression, zero-inflated Poisson (ZIP) regression, and NB regression. Maximum likelihood (ML) method is used to estimate the unknown parameters of these models. Two other feasible estimators for estimating the dispersion parameter in the NB regression model are also examined: a moment estimator and a regression-based estimator. These models and estimators are evaluated based on their (i) estimated regression parameters, (ii) overall goodness-of-fit, (iii) estimated relative frequency of truck accident involvements across road sections, (iv) sensitivity to the inclusion of short road sections, and (v) estimated total number of truck accident involvements. Data from the Highway Safety Information System are employed to examine the performance of these models in developing such relationships. The evaluation results suggest that the NB regression model estimated using the moment and regression-based methods should be used with caution. Also, under the ML method, the estimated regression parameters from all three models are quite consistent and no particular model outperforms the other two models in terms of the estimated relative frequencies of truck accident involvements across road sections. It is recommended that the Poisson regression model be used as an intial model for developing the relationship. If the overdispersion of accident data is found to be moderate or high, both the NB and ZIP regression models could be explored. Overall, the ZIP regression model appears to be a serious candidate model when data exhibit excess zeros, e.g. due to underreporting. However, the interpretation of the ZIP model can be difficult.

Keywords—Truck accidents, Geometric design, Poisson, Zero-inflated Poisson, Negative binomial

1. INTRODUCTION

The relationships between vehicle accidents and geometric design of road sections, such as horizontal curvature, vertical grade, lane width, and shoulder width, have been studied using multiple linear regression models in numerous previous studies (Roy Jorgensen Associates, Inc. 1978; Zegeer et al. 1987; Okamoto and Koshi 1989; Zegeer et al. 1990). Because the occurrences of vehicle accidents are typically sporadic across the road network, in most vehicle accidents-geometric design studies the analysts are faced with a problem of dealing with a large number of road sections that had no reported accidents during the observed period. For example, in a study by Zegeer et al. (1990), 55.7% of the road sections they studied had had no reported vehicle accidents in a five-year period, and in another study by Miaou and Lum (1993b), over 80% of the road sections had no reported truck accidents during a one-year period. This suggests that for a period of several years most of the road sections considered would have a high probability of being observed with no accidents. In other words, the underlying distribution of the occurrences of vehicle accidents on most of the road sections during the observed period is positively or rightly skewed. It has been demonstrated that the conventional multiple linear regression models, which rely on normal assumption, lack the distributional property necessary to describe adequately the random and discrete vehicle accident events on the road (Jovanis and Chang 1986; Saccomanno and Buyco 1988; Miaou and Lum 1993a). As a result, these linear regression models are not appropriate to make probabilistic statements about vehicle accidents on the road, and test statistics derived from these models are questionable.

The unsatisfactory property of linear regression models has led to the investigation of the Poisson and negative binomial (NB) regression models in recent studies (Maycock and Hall 1984; Joshua and Garber 1990; Miaou et al. 1991; Miaou et al. 1992; Miaou et al. 1993; Miaou and Lum 1993b). Although the Poisson and NB regression models have been found to have desirable distributional property to describe the vehicle accidents-geometric design relationship, these models are not without limitations. In addition, the relative performance of these models in establishing such relationships has not been fully evaluated.

The objective of this paper is to evaluate the statistical performance of the Poisson and NB regression models in establishing the relationship between truck accidents and geometric design of road sections. Three types of models are considered: Poisson regression, zero-inflated Poisson (ZIP) regression, and NB regression. Maximum likelihood (ML) method is used to estimate the unknown parameters of these models. Two other feasible estimators for estimating the dispersion parameter in the NB regression model are also examined: a moment estimator and a regression-based estimator. These models and estimators are evaluated based on their (i) estimated regression parameters and associated t-statistics, (ii) overall goodness-of-fit, (iii) estimated relative frequency of truck accident involvements across road sections, (iv) sensitivity to the inclusion of short road sections, and (v) estimated total number of truck accident involvements. Data from the Highway Safety Information System (HSIS), a highway safety data base administered by the Federal Highway Administration (FHWA), are employed to examine the performance of these models in developing such relationships.

The remaining paper is organized as follows. First, three types of Poisson and NB regression models used for studying truck accidents and geometric design relationships are presented. Second, the performance of these models are evaluated using the HSIS data. The last section concludes the study.

2. MODELS AND ESTIMATORS

Consider a set of n road sections of a particular roadway type, say, rural Interstate. Let Y_i be a random variable representing the number of trucks involved in accidents on road section i during a period of one year, where $i = 1, 2, \ldots, n$. Note that the same road section in different sample periods are considered as separate road sections; this allows the year-to-year changes in geometric design, traffic conditions, and other relevant attributes to be considered in the model. Further, let the actual observation of Y_i during the period be denoted as y_i, where $y_i = 0, 1, 2, 3, \ldots$ and $i = 1, 2, \ldots, n$. Also, let the amount of truck travel (or truck exposure) during the sample year on this road section be v_i and computed as $365 \times \text{AADT}_i \times \text{T\%}_i \times \ell_i$, where AADT_i is the average annual daily traffic (in number of vehicles), $\text{T\%}_i$ is the percentage of trucks in the traffic stream, and ℓ_i is the length of road section i. Associated with each road section i, there is a $k \times 1$ covariate vector, $\mathbf{x}_i$ describing its geometric characteristics, traffic conditions, and other relevant attributes. The transpose of the covariate vector is denoted by $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, \mathbf{x}_{ik})$. Without loss of generality, let the first covariate x_{i1} be a dummy variable equal to one for all i (i.e. $x_{i1} = 1$). Some of the covariates can be 0, 1 dummy variables, indicating the presence or absence of a condition.

The general forms of these three types of regression models and brief descriptions of their estimation procedures are presented in this section. All models are formulated under the assumption that (i) the occurrences of truck accidents on different road sections are independent, and (ii) truck miles data and other covariates are free from errors.

Poisson regression model

The Poisson regression model considered in this study is the one used by Miaou et al. (1991):

$$p(Y_i = y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i = 1, 2, 3, \ldots, n. \tag{1}$$

where

$$\mu_i = E(Y_i) = \nu_i[e^{\mathbf{x}_i'\boldsymbol{\beta}}] = \nu_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \qquad i = 1, 2, 3, \ldots, n. \tag{2}$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown regression parameters, the transpose of which is denoted by $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_k)$. This model assumes that Y_i, $i = 1, 2, \ldots, n$, are independently and Poisson distributed with mean μ_i. The expected number of trucks involved in accidents μ_i or $E(Y_i)$ in this model is proportional to truck travel ν_i. The model also assumes an exponential rate function, $\lambda_i = E(Y_i)/\nu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, which ensures that accident-involvement rate is always nonnegative. This type of rate function has been widely employed in statistical literature and found to be very flexible in fitting different types of count data (e.g. Cox and Lewis 1966; Cameron and Trivedi 1986; Frome, Cragle, and McLain 1990). Note that whenever appropriate higher order and interaction terms of covariates can be included in eq. (2) without difficulties.

The regression parameters $\boldsymbol{\beta}$ of this model can be estimated using the ML method (Cramer 1986), the quasi-likelihood method (McCullagh and Nelder 1983), or the generalized least squares method (Carroll and Ruppert 1989). The estimated parameters from the last two methods would converge to those from the ML method as more iterations are used. For deriving the asymptotic variance and t-statistics of the estimated parameters using the second derivative of the loglikelihood function, see Cramer (1986) for reference.

A limitation of using the Poisson regression model, which is well known in the statistical literature (e.g. Cox 1983; Dean and Lawless 1989), is that the variance of the data is restrained to be equal to the mean, $\mathrm{Var}(Y_i) = E(Y_i) = \mu_i$. In many applications, count data were found to display extra variation or overdispersion relative to a Poisson model (e.g. Dean and Lawless 1989). That is, the variance of the data was greater than what the Poisson model indicated. In vehicle accidents-geometric design studies, the overdispersion could come from several possible sources, e.g. omitted variables, uncertainty in exposure data and covariates, and nonhomogeneous highway environment (Miaou et al. 1993).

The consequences of ignoring the extra variations in the Poisson regression models are that consistent estimates, such as the ML estimates (MLE), of the regression parameters under the Poisson model are still consistent; however, the variances of the estimated parameters would tend to be underestimated. In other words, when the sample size n is large, the MLE $\hat{\boldsymbol{\beta}}$ under the Poisson regression model would still be close to the true parameter $\boldsymbol{\beta}$, but we may overstate the significance levels of the estimated parameters (Cameron and Trivedi 1990). (Note that this is under the assumption that eq. (2) is correctly specified.)

To correct for the overdispersion problem for the Poisson model, Wedderburn (1974) suggested that one could assume that the variance of Y_i is $\tau\mu_i$ instead of μ_i, where τ is referred to as overdispersion parameter (and $\tau \geq 1$). It was also suggested that the overdispersion parameter τ could be estimated by $X^2/(n - k)$, where X^2 is the Pearson's chi-square statistic, n is the number of observations (i.e. the number of road sections), and k is the number of unknown regression parameters in the Poisson model. The Pearson's X^2 statistic is computed as $\Sigma_i(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$, where $\hat{\mu}_i = \nu_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$. A better estimate of the asymptotic t-statistic for each regression parameter is $\tau^{-1/2}$ times that obtained from the original Poisson regression model based on the ML method (Agresti 1990).

Zero-inflated Poisson regression model

The particular ZIP regression model considered in this study has the following form:

$$\begin{aligned} p(Y_i = y_i) &= e^{-\theta r_i} && \text{if } y_i = 0 \\ &= \left(\frac{1 - e^{-\theta r_i}}{1 - e^{-r_i}}\right)\frac{r_i^{y_i} e^{-r_i}}{y_i!} && \text{if } y_i = 1, 2, 3, \ldots \end{aligned} \quad (3)$$

and

$$r_i = \nu_i[e^{\mathbf{x}_i'\boldsymbol{\beta}}] = \nu_i[e^{\Sigma_{j=1}^{k} x_{ij}\beta_j}] \quad i = 1, 2, 3, \ldots, n. \quad (4)$$

where $0 < \theta \leq 1$. (Note that for $\theta > 1$, the probability of observing zeros is deflated rather than inflated.) Under this ZIP regression model, the mean and variance of Y_i can be shown to be

$$\mu_i = E(Y_i) = \left(\frac{1 - e^{-\theta r_i}}{1 - e^{-r_i}}\right) r_i \quad \text{and}$$

$$\mathrm{Var}(Y_i) = \mu_i + \left(\frac{1 - e^{r_i(\theta - 1)}}{e^{\theta r_i} - 1}\right)\mu_i^2 = \mu_i + \phi_i\mu_i^2 \quad (5)$$

where ϕ_i is a function of r_i and θ. When $\theta = 1$, the ZIP regression model is identical to the Poisson regression model presented in eqs. (1) and (2). Also, when $0 < \theta < 1$, one can show that the variance of Y_i exceeds its mean. Thus, the ZIP regression model allows overdispersion in the data due to excess zeros when compared to the classic Poisson regression model.

One reason for considering the ZIP regression model is the potential underreporting of vehicle accidents, especially minor injury and property-damage accidents. If the number of trucks involved in accidents on road sections follow a Poisson distribution, then, because of underreporting, the "reported" number of trucks involved in accidents would follow a ZIP distribution under some underreporting conditions.

Although the ZIP regression is more flexible than the Poisson regression, the interpretation of the ZIP regression can be difficult. For example, the expected number of trucks involved in accidents μ_i is related to truck travel ν_i and other covariates $\mathbf{x}_i'$ in a much more complicated way in the ZIP regression than that in the Poisson regression, and it is not as easy to see how an increase in ν_i or $\mathbf{x}_i'$ would increase or decrease the mean μ_i in the ZIP regression. Note, however, that the ZIP regression model still ensures that truck accident involvement rate is always nonnegative.

The ML estimation and other possible variations of the ZIP regression models have been discussed in Lambert (1992). In this study, the ML method is used to estimate the unknown parameters $\boldsymbol{\beta}$ and θ, and their asymptotic standard deviations and t-statistics are estimated using the second derivative of the loglikelihood function (or the observed Fisher information matrix).

Negative binomial regression model

To deal with the overdispersion problem in count data, one commonly used distribution is the NB distribution. The NB regression model considered in this study has the following form:

$$p(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \quad y_i = 0, 1, 2, \ldots \tag{6}$$

where

$$\mu_i = E(Y_i) = \nu_i[e^{x_i'\beta}] = \nu_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \qquad i = 1, 2, 3, \ldots, n. \tag{7}$$

and the variance of Y_i is

$$\mathrm{Var}(Y_i) = \mu_i + \alpha\mu_i^2 \tag{8}$$

where $\alpha \geq 0$ and is usually referred to as dispersion parameter. From eq. (8) one can see that this model allows the variance to exceed the mean. Also, the Poisson regression model can be regarded as a limiting model of the negative binomial regression model as α approaches 0.

The ML estimation of the NB regression model and the calculation of associated statistics are described in detail by Lawless (1987). The moment estimation, which was first suggested by Breslow (1984), is also commonly used for estimating the parameters in the NB model. For comparison, in this study the moment estimation method as described in Lawless (1987) is also used for parameter estimation. The method is an iterative procedure which iterates until the estimated dispersion parameter converges.

Another estimation method considered in this study for estimating the NB regression model is a regression-based estimation method suggested by Cameron and Trivedi (1986, 1990). The method is an iterative estimation procedure described below.

Step 1. Give an initial estimate of α, say, $\hat{\alpha}_1$.

Step 2. Estimate $\boldsymbol{\beta}$ using the ML method, assuming $\alpha = \hat{\alpha}_1$. Let the estimate of $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}$.

Step 3. Obtain the following regression estimator for α:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \hat{\mu}_i^2[(y_i - \hat{\mu}_i)^2 - \hat{\mu}]}{\sum_{i=1}^{n} \hat{\mu}_i^4} \tag{9}$$

where $\hat{\mu}_i = \nu_i\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$.

Step. 4. If $|\hat{\alpha} - \hat{\alpha}_1| < \varepsilon$, stop; else let $\hat{\alpha}_1 = \hat{\alpha}$ and go to Step 2. Here, ε is a small positive number equal to, e.g. 0.001.

The NB regression model using the ML method was used in Miaou et al. (1993) to establish truck accidents-geometric design relationships for three different roadway classes. Although the NB regression model is more general than the Poisson regression model, it requires more extensive computations to estimate model parameters and to generate inferential statistics than the Poisson regression model. Furthermore, the statistical property of different estimators, e.g. the MLE and moment estimators, of the NB regression model under different sample sizes have not yet been fully investigated (Lawless 1987).

Although the discussion above does not distinguish accidents by truck configuration and accident severity, in principal, these three types of Poisson and NB regression models could be applied to any truck type and accident-severity type of interest, provided that there are enough accident data and that truck exposure by truck type can be properly estimated.

3. MODEL EVALUATIONS

Data source

Data from the HSIS are employed to evaluate the performance of the Poisson and NB regression models (and estimators) described above in terms of its ability to establish truck accidents and geometric design relationships. Specifically, accidents involving large trucks on rural interstate highways from Utah are used. Note that here large trucks are de-

Table 1. Variable definitions and summary statistics of the 8,263 rural interstate road sections

Variable	Notation & Definition (for section *i*)	Min	Max	Mean	% Zero
Number of Trucks Involved in Accidents	y_i	0	8	0.20	86
Section Length (in mi)	ℓ_i	0.01	7.77	0.45	0
Truck Miles or Truck Exposure (in 10^6 truck-miles)	v_i = [365×AADT$_i$×(T%$_i$/100)×ℓ_i]/10^6, where T%$_i$ is percent trucks (366 for leap years).	8×10^{-4}	5.03	0.25	0
Dummy Intercept	$x_{i1} = 1$				
Dummy Variable for Year 1986, representing year-to-year changes due to random fluctuations, annual trend, and omitted variables such as weather.	x_{i2} = 1, if the road section is in year 1986 = 0, otherwise				
Dummy Variable for Year 1987 (See above explanation)	x_{i3} = 1, if the section is in 1987 = 0, otherwise				
Dummy Variable for Year 1988 (See above explanation)	x_{i4} = 1, if the section is in 1988 = 0, otherwise				
Dummy Variable for Year 1989 (See above explanation)	x_{i5} = 1, if the section is in 1989 = 0, otherwise				
AADT per Lane (in 1000's of vehicles), a surrogate variable to indicate traffic conditions or traffic density.	x_{i6} = (AADT$_i$/number of lanes$_i$)/1000	0.35	12.04	1.80	0
Horizontal Curvature, HC, (in degrees per 100-ft arc)	x_{i7}	0	12.00	1.00	67
Length of Original Horizontal Curve, LHC, (in mi) from which this curve was subdivided for creating homogeneous sections; only for HC > 1 and LHC ≤ 1.	x_{i8} = LHC, if $x_{i7}>1$ and LHC ≤ 1 mi. = 1.0, if $x_{i7}>1$ and LHC > 1 mi. = 0, if $x_{i7}\leq 1$	0	0.96	0.05	81
Vertical Grade, VG, (in percent)	x_{i9}	0	8.00	2.14	20
Length of Original Vertical Grade, LVG, (in mi) from which this section was subdivided for creating homogeneous sections; only for sections with VG > 2 and LVG ≤ 2.	$x_{i,10}$ = LVG, if $x_{i9}>2$ and LVG ≤ 2 mi. = 2.0, if $x_{i9}>2$ and LVG > 2 mi. = 0, if $x_{i9}\leq 2$	0	2.00	0.21	74
Deviation of Paved Inside Shoulder Width (per direction) from an "ideal" width of 12 ft (3.66 m).	$x_{i,11}$ = max{0, 12 - paved inside shoulder width}	4.00	12.00	8.16	0
Percent Trucks in the traffic stream (e.g., 15)	$x_{i,12}$	7.00	57.00	24.13	0
HC × LHC	$x_{i,13} = x_{i7} \times x_{i8}$	0	2.88	0.18	81
VG × LVG	$x_{i,14} = x_{i9} \times x_{i,10}$	0	13.37	0.97	74

(1) All of the sections are 12 ft (3.66 m) in lane width. (2) About 89% of the sections have 4 lanes; and all sections have paved outside shoulder width of 10 ft (3.05 m). (3) Total number of trucks involved in accidents = 1,643; total highway lane-mi = 14,731; and total truck travel = 2,030×10^6 truck mi (3 248×10^6 truck km). (4) 1 mi = 1.61 km; 1 ft = 0.3048 m.

fined as trucks with gross vehicle weight rating of 10,000 lb. or over. Among the five HSIS states available, Utah was considered to be the state that had the most complete information on highway geometric design (Miaou et al. 1993). In addition, this particular state was the only HSIS state with a "historical" road-inventory file in which year-to-year changes on highway geometric design and traffic conditions were recorded. Thus, accidents in a given year could be matched to the road-inventory information of the same period. For this study, truck-accident and road-inventory data from 1985 to 1989 are used.

Accidents, characteristics of road sections, and covariates

Detailed descriptions of the data can be found in Miaou et al. (1993). The time period considered in this study is one year, which means that the same road section, even if nothing had changed, is considered as five independent sections—one for each year from 1985 to 1989. As indicated earlier, this allows the year-to-year changes on highway geometric design and traffic conditions to be considered in the model. There is a total of 8,263 homogeneous road sections during the five-year period. These road sections constitute 14,731 lane-miles of roadway, which covers over 98% of the total rural interstate lane-miles in Utah during the period. Data for each year contain roughly one-fifth of the total sections and lane-miles. The section lengths vary from 0.01 to 7.77 mi (0.016 to 12.43 km)—with an average of 0.45 mi (0.72 km). Descriptive statistics of these 8,263 road sections on truck accident involvements and truck miles traveled are given in Table 1.

During the five-year period, 1,643 large trucks reported to be involved in accidents on these highway sections, regardless of truck configuration and accident severity type. With the total truck miles estimated to be 2,030 million (3,248 million truck kilometers), the overall truck accident involvement rate was 0.81 truck accident involvements per million truck miles (0.51 truck accident involvements per million truck kilometers). These accidents occurred on only 14% of the 8,263 road sections. The maximum number of trucks involved in accidents on an individual road section was eight. On average, each section had 0.20 trucks involved in accidents in one year.

The covariates considered for individual road sections and their definitions are also presented in Table 1. They include (i) yearly dummy variables to capture year-to-year changes in the overall truck accident involvement rate due, e.g. to long-term trend, annual random fluctuations, speed limit change, and changes in omitted variables such as weather; (ii) AADT per lane, used as a surrogate measure for traffic density; (iii) horizontal curvature; (iv) vertical grade; and (v) deviation of paved inside (or left) shoulder width from an "ideal" width of 12 ft per direction. Because all of the road sections are 12 ft in lane width, about 89% of them have four lanes, and all road sections have paved outside (or right) shoulder width of 10 ft per direction, we are unable to test the effects of these variables.

It has been suggested that as length of grade increases to a point that can slow a truck to a speed significantly slower than the speed of the traffic stream (e.g. 10 mi/h or 16 km/h), the accident rate increases (Roy Jorgensen Associates 1978). Also, for a fixed curvature degree, as the length of curve increases, the accident rate increases (Zegeer et al. 1990). In order to test the effects of length of curve and length of grade on truck accident involvement rate, two covariates—length of original curve and length of original grade—are considered. Because each curve or grade considered in the model may have been subdivided from a longer curve or grade for achieving total homogeneity, for each road section in the model these two covariates are defined as the length of the original undivided curve or grade to which this section belongs. In addition, these two covariates are defined only for curves with horizontal curvatures greater than 1 degree and sections with grade greater than 2%. (Note that these two covariates are set equal to 0 if horizontal curvature is less than or equal to 1 degree, or if vertical grade is less than or equal to 2%.) This definition is based on an assumption that the length of a mild curve or grade has no aggravated effect on truck accident involvements. The interactions of horizontal curvature and length of original curve, vertical grade and length of original grade, and horizontal curvature and vertical grade are also considered.

Percentage of trucks in the traffic stream is included in the model to evaluate the effects of car-truck mix. Previous studies suggested that as the percentage of trucks increases, truck accident involvement rate decreases. One possible reason is that, *for a constant vehicle density*, as the percentage trucks increases, the frequency of lane changing and overtaking movements by cars decreases. Also, previous records showed that more trucks are involved in truck-car multivehicle accidents than in truck-truck accidents (Jovanis and Chang 1986).

Model estimation results

The estimated parameters and associated t-statistics of five Poisson and NB regression models and estimators described in the previous section are

Table 2. Estimated regression parameters and associated statistics

	Poisson	ZIP	NB-ML	NB-MOM	NB-REG	Poisson	ZIP	NB-ML	NB-MOM	NB-REG
	8,263 Sections (with section length ≥ 0.01 mi)					7,004 Sections (with section length > 0.05 mi)				
β_1 Dummy intercept	-0.43176 (-1.20)	-0.09436 (-0.32)	-0.26521 (-0.76)	0.09498 (0.21)	-0.36534 (-1.22)	-0.52610 (-1.57)	-0.22742 (-0.76)	-0.40655 (-1.16)	-0.38360 (-1.03)	-0.53001 (-1.76)
β_2 Dummy variable for 1986	-0.18385 (-1.71)	-0.17790 (-2.10)	-0.20439 (-1.96)	-0.22171 (-1.60)	-0.18583 (-2.08)	-0.17176 (-1.71)	-0.16795 (-1.95)	-0.19197 (-1.82)	-0.19135 (-1.70)	-0.18545 (-2.04)
β_3 Dummy variable for 1987	-0.16146 (-1.52)	-0.16549 (-1.98)	-0.13961 (-1.35)	-0.09544 (-0.69)	-0.15523 (-1.75)	-0.16087 (-1.62)	-0.16982 (-2.00)	-0.14906 (-1.42)	-0.13412 (-1.19)	-0.16417 (-1.82)
β_4 Dummy variable for 1988	-0.11151 (-1.05)	-0.12163 (-1.45)	-0.08400 (-0.80)	-0.03912 (-0.28)	-0.10010 (-1.13)	-0.09624 (-0.97)	-0.11207 (-1.32)	-0.06296 (-0.59)	-0.05816 (-0.51)	-0.10081 (-1.12)
β_5 Dummy variable for 1989	-0.31116 (-2.83)	-0.32162 (-3.71)	-0.31145 (-2.90)	-0.28615 (-2.00)	-0.30980 (-3.38)	-0.29970 (-2.92)	-0.31216 (-3.56)	-0.29360 (-2.69)	-0.28684 (-2.46)	-0.30393 (-3.27)
β_6 AADT per lane (10^3)	0.02440 (1.27)	0.00669 (0.44)	0.02462 (1.22)	0.01657 (0.57)	0.02196 (1.35)	0.02522 (1.42)	0.00831 (0.54)	0.02608 (1.28)	0.02683 (1.20)	0.02599 (1.58)
β_7 Horizontal curvature	0.08886 (2.51)	0.11728 (4.20)	0.07365 (2.31)	0.04617 (1.17)	0.07988 (2.76)	0.09617 (2.84)	0.12403 (4.30)	0.08031 (2.42)	0.07549 (2.17)	0.09102 (3.04)
β_{13} (Horizontal curvature)× (Length of original curve)	0.23421 (2.22)	0.20988 (2.51)	0.27707 (2.77)	0.35943 (2.68)	0.26460 (3.07)	0.22188 (2.23)	0.20057 (2.33)	0.27211 (2.64)	0.28944 (2.63)	0.23864 (2.67)
β_9 Vertical grade	0.07782 (2.25)	0.07713 (2.76)	0.08678 (2.72)	0.10666 (2.61)	0.07461 (2.63)	0.07822 (2.42)	0.07625 (2.69)	0.08488 (2.62)	0.09121 (2.67)	0.07920 (2.73)
β_{14} (Vertical grade)×(Length of original grade)	0.03397 (1.81)	0.02398 (1.63)	0.02790 (1.45)	0.01470 (0.57)	0.03606 (2.27)	0.03109 (1.77)	0.02146 (1.43)	0.02452 (1.25)	0.02021 (0.96)	0.03019 (1.86)
β_{11} Deviation of paved inside shoulder width from 12 ft	0.08576 (1.90)	0.10207 (2.74)	0.07092 (1.61)	0.03369 (0.59)	0.08128 (2.15)	0.09481 (2.26)	0.11615 (3.10)	0.08589 (1.94)	0.08174 (1.74)	0.09709 (2.56)
β_{12} Percent trucks (e.g., 15)	-0.02523 (-4.70)	-0.02707 (-6.39)	-0.02653 (-4.96)	-0.02860 (-4.01)	-0.02586 (-5.75)	-0.02531 (-5.06)	-0.02684 (-6.26)	-0.02677 (-4.95)	-0.02672 (-4.61)	-0.02558 (-5.60)
τ	1.57					1.32				
θ		0.58738 (19.08)					0.58217 (18.90)			
α			0.94652 (8.89)	4.3322	0.1286			0.94128 (8.77)	1.4595	0.1352
$L(\hat{\beta})$	-3771.0	-3723.6	-3682.4	-3840.2*	-3738.6*	-3589.1	-3540.9	-3504.8	-3514.7*	-3556.3*
AIC Value	7566.0	7473.2	7390.7	7704.3*	7501.2*	7202.3	7107.7	7035.6	7053.4*	7136.6*
Expected vs Observed Total Truck Accident Involvements	1,644.3 1,643.0	1,657.3 1,643.0	1,702.6 1,643.0	1,773.3 1,643.0	1,658.3 1,643.0	1,604.5 1,603.0	1,615.7 1,603.0	1,652.5 1,603.0	1,664.5 1,603.0	1,613.4 1,603.0

Notes: (1) Values in parentheses are (adjusted) asymptotic t-statistics of the parameters above.
(2) ----- Not included in the model.
(3) * Computed using the negative binomial model.

Table 3. Observed versus estimated relative frequencies of truck accident involvements across the 8,263 road sections

Number of Truck Accident Involvements (k)	Observed Relative Frequency of Road Sections with k Truck Accident Involvements (f_k, in percent)	Estimated Relative Frequency of Road Sections with k Truck Accident Involvements ($\hat{f}_k$, in percent)		
		Poisson	ZIP	NB-ML
$k=0$	86.1915	84.7731 (-1.4184)	86.4029 (0.2114)	86.2758 (0.0843)
$k=1$	9.7180	11.9099 (2.1919)	9.5757 (-0.1423)	9.9112 (0.1932)
$k=2$	2.7351	2.4039 (-0.3312)	2.5832 (-0.1519)	2.3406 (-0.3945)
$k=3$	0.9803	0.6241 (-0.3562)	0.8629 (-0.1174)	0.7880 (-0.1923)
$k=4$	0.2178	0.1914 (-0.0264)	0.3324 (0.1146)	0.3272 (0.1094)
$k \geq 5$	0.1573	0.0976 (-0.0597)	0.2429 (0.0856)	0.3572 (0.1999)
$k \geq 0$	100	100 (0.0)	100 (0.0)	100 (0.0)

Values in parentheses are (estimated relative frequencies - observed relative frequencies) in percent.

presented in Table 2. Discussion on variable selection and testing procedures can be found in Miaou and Lum (1993b). The Poisson and ZIP regression models are estimated using the ML method. The NB regression model is estimated using the ML method (NB-ML), moment method (NB-MOM), and regression-based method (NB-REG). The loglikelihood function evaluated at the estimated parameters, L($\hat{\boldsymbol{\beta}}$), and the Akaike Information Criterion (AIC) value (see, e.g. Bozdogan, 1987) for each model are also given in the table. Note that AIC $= -2\text{L}(\hat{\boldsymbol{\beta}}) + 2k$, where k is the total number of unknown regression parameters in the model. Estimated models with high loglikelihood function and low AIC values are preferred. The Wedderburn's overdispersion parameter τ for the Poisson regression model is computed to adjust the estimated t-statistics from the MLE. Furthermore, the expected total number of trucks involved in accidents across road sections from the model ($\Sigma_i \hat{\mu}_i = \Sigma_i \nu_i \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}})$) is compared with the observed total ($\Sigma_i y_i$).

In order to examine the effect of short road sections on the estimation of different models and estimators, we remove road sections with section length less than or equal to 0.05 mi (0.08 km). As a result, a total of 1,259 road sections are eliminated. The remaining 7,004 road sections, which had 1,603 truck accident involvements, are used to develop new models. The results are also presented in Table 2.

To evaluate how well the Poisson regression, ZIP regression, and NB-ML models estimate the relative frequencies of truck accident involvements across road sections, the following comparison is made. First, the relative frequency of road sections with k trucks involved in accidents is computed and denoted by f_k. More specifically, f_k is the percentage of road sections which had k truck accident involvements during the sampled year (i.e. f_k = number of road sections with k truck accident involvements/total number of road section n). Second, the estimated relative frequency is computed as: $\hat{f}_k = \Sigma_i \hat{p}(Y_i = k)/n$, where $\hat{p}(Y_i = k)$ is the estimated probability of having k trucks involved in accidents, given the estimated parameters. The observed and estimated relative frequencies (f_k and $\hat{f}_k$) and their differences ($\hat{f}_k - f_k$) are presented in Table 3. In this table, the relative frequencies are computed for $k = 0, 1, 2, 3, 4$ and for $k \geq 5$. (Note that this kind of comparison has been used by Lambert [1992].)

Model performance

From Tables 2 and 3, the following observations can be made:

1. Estimated regression parameters and associated t-statistics:
The comparison of the estimated parameters and t-statistics in Table 2 for the three models under the ML estimation method suggests not only that the conclusions reached regarding the significance level of the relationships between truck accidents and the examined traffic and geometric design variables are quite consistent, but also that most of the estimated regression parameter values are quite close. In addition, all of the estimated regression parameter values related to traffic and geometric design variables have the expected algebraic sign. Another observation is that the ZIP regression model seems to suggest slightly stronger relationships between truck accidents and horizontal curvature and between truck accidents and paved inside shoulder width than the other two models.
2. Sensitivity to the inclusion of short road sections:

Table 4. Truck accident involvement rates and accident probabilities of three example rural interstate road sections

Variables	Model	Section #1	Section #2	Section #3
Year		1989	1989	1989
Lane Width (ft)		12	12	12
Number of Lanes		4	4	4
Section Length (mi)		0.3	0.3	0.3
AADT (number of vehicles)		5,000	25,000	50,000
Horizontal Curvature (degree/100 ft arc)		0	3	6
Length of Original Curve (mi)		0	0.5	0.5
Vertical Grade (percent)		0	3	3
Length of Original Grade (mi)		0	0.3	0.3
Paved Inside Shoulder Width (ft)		10	6	6
Paved Outside Shoulder Width (ft)		10	10	10
Percent Trucks		25	25	25
Truck Accident Involvement Rate (λ_i)	Poisson ZIP NB-ML	0.3089 0.2464 0.3439	1.1881 1.1554 1.2989	2.5669 3.0861 2.8631
Expected Number of Truck Accident Involvements ($E(Y_i)$ or μ_i)	Poisson ZIP NB-ML	0.0423 0.0337 0.0471	0.8131 0.7908 0.8889	3.5135 4.2241 3.9189
Variance of Truck Accident Involvements ($Var(Y_i)$)	Poisson ZIP NB-ML	0.0423 0.0345 0.0492	0.8131 1.0410 1.6368	3.5135 5.3787 18.4556
Probability of Truck Accident Involvements ($p(Y_i)$)		Fig. 1(a)	Fig. 1(b)	Fig. 1(c)

1 mi = 1.61 km; 1 ft = 0.3048 m.

Except the NB-MOM model, the estimated regression parameter values for each model only change slightly after the short road sections (with $\ell_i \leq 0.05$ mi) are removed from the data. Some of the estimated regression parameters of the NB-MOM model are somewhat different from those of the NB-ML model when short road sections are included in the model estimation (e.g. β_7, β_{14}, and β_{11}). However, after removing the short road sections, the estimated regression parameter values from the two models become very close. This indicates that the NB regression model using the moment estimation is sensitive to the inclusion of short road sections. (Note that it has been shown by Miaou and Lum [1993] that those vehicle accidents-geometric design models based on conventional linear regressions are also very sensitive to the inclusion of short road sections.)

3. Overdispersion and dispersion parameters: When all road sections are considered in the model, the Wedderburn's overdispersion parameter τ for the Poisson regression model is 1.57. This suggests that the truck-accident data used in this study are moderately overdispersed, i.e. moderately higher than what the Poisson model has implied. Also, when short road sections are excluded from the model, the Wedderburn's overdispersion parameter decreases to 1.32, indicating that the overdispersion measure is somewhat inflated when short road sections are included in the model. For the NB regression model, the estimated dispersion parameter from the three estimators are quite different. The inclusion of short road sections does not affect the estimation from the ML and regression-based estimators, but tends to inflate the dispersion parameter when estimated using the

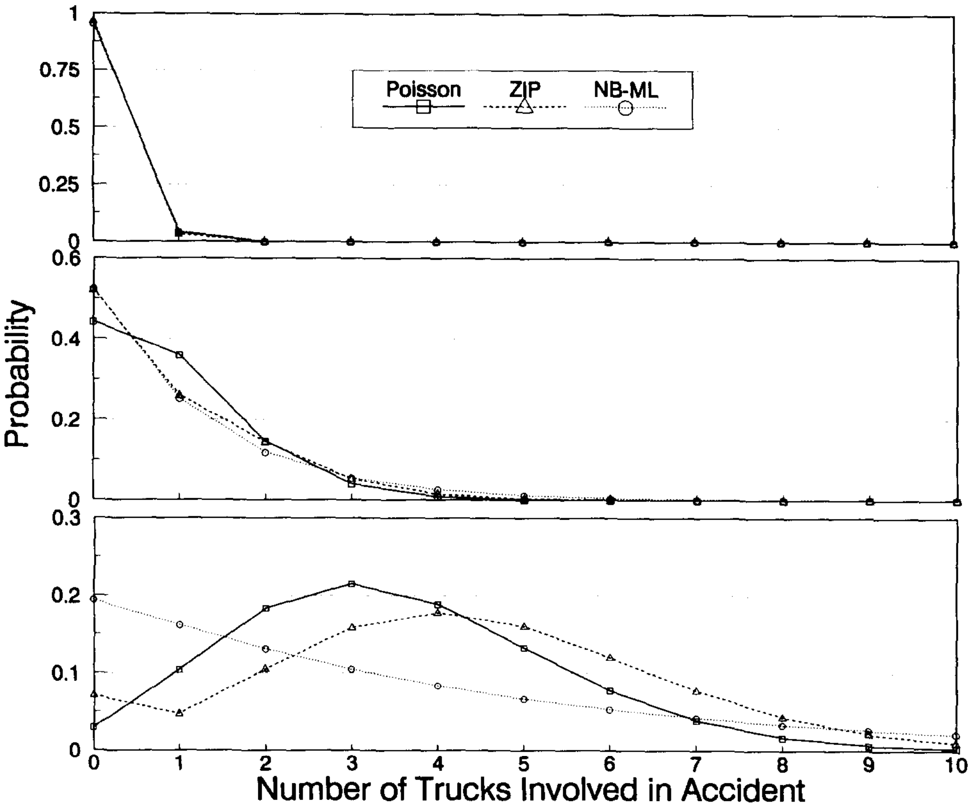

Fig. 1. Example probability distributions of truck accident involvements.

moment method. Also, the regression estimator seems to have underestimated the dispersion parameter when compared to the ML estimator, Poisson regression model, and ZIP regression model.

4. AIC value, loglikelihood function, and relative frequency:
   Based on the loglikelihood function and the AIC value, the NB-ML model has better performance than the Poisson and ZIP regression models. However, by examining the estimated relative frequencies in Table 3, it is found that no particular model is performing better than the other two models for all ks. For $k = 0$, the NB-ML model performs the best and the ZIP model performs quite well; for $k = 1, 2, 3$, the ZIP model performs the best; and for $k \geq 4$, the Poisson model performs the best. Therefore, it seems that the NB-ML model is performing better under the loglikelihood function and the AIC value simply because it provides better estimated probabilities of having no reported truck accident for different road sections and over 86% of the road sections were reported to have no truck accident.
5. Expected versus observed total truck accident involvements:
   Except the NB-MOM model, the expected total truck accident involvements are reasonably close to (or specifically within 4% of) the observed total when all road sections are considered in the model. After removing short road sections, the expected total from the NB-MOM model is found to be closer to the observed total. This indicates that the estimated overall accident involvement rate is inflated in the NB-MOM model when short road sections are included. (Note that it was demonstrated in Miaou and Lum [1993] that conventional linear regression models may seriously overestimate or underestimate the total number of accident involvements across road sections.)

Examples of accident probability

To see how accident probability distribution varies under the three regression models as the expected number of truck accident involvements increases, three hypothetical road sections are created to represent a low, a medium, and a high truck accident involvement road section and are given in Table 4. The estimated accident probability distributions of the three hypothetical road sections are given in Fig. 1. When the expected number of truck accident involvements ($E(Y_i)$ is low (say, < 1), the estimated probability distributions ($\hat{p}(Y_i)$) from the three models are quite similar in shape and magnitude. When $E(Y_i)$ is high (say, > 2), $\hat{p}(Y_i)$ from the three models deviate from one another substantially in shape and magnitude. Particularly, the NB-ML model suggests a much more diffused accident probability distribution than the other two models as $E(Y_i)$ increases beyond 2.

4. CONCLUSIONS

The performance of the Poisson and NB regression models in establishing the relationship between truck accidents and geometric design of road sections is evaluated using the HSIS data. The Poisson and ZIP regression models are estimated using the ML method, while the NB regression model are estimated using the ML, moment, and regression-based estimators. The NB model based on the moment method is quite sensitive to the inclusion of short road sections. While the NB model using the regression estimator tends to understate the dispersion of the data. Both estimators should be used with caution. Under the ML method, the estimated regression parameters from all three models are quite consistent and no particular model outperforms the other two models in terms of the estimated relative frequencies of truck accident involvements across road sections. The NB model performs the best in estimating the frequency of road sections with zero truck accident involvement. The ZIP model, on the other hand, performs the best in estimating the frequencies of road sections with one, two, and three truck accident involvements. While the Poisson model performs the best in estimating the frequencies of road sections with four or more truck accident involvements. It is recommended that the Poisson regression model be used as an initial model for developing the relationship. If the overdispersion of accident data is found to be moderate or high (e.g. when the Wedderburn's overdispersion parameter > 1.3), both the NB and ZIP regression models could be explored. Overall, the ZIP regression model appears to be a serious candidate model for studying the relationships when accident data exhibit excess zeros due, e.g. to underreporting. However, the interpretation of the ZIP model can be difficult.

Acknowledgements—This paper is based on the results of a research project sponsored by the Office of Safety and Traffic Operations R&D, Federal Highway Administration (FHWA), the U.S. Department of Transportation. The opinions expressed in this paper are, however, solely those of the author.

REFERENCES

Agresti, A. Categorical data analysis. New York: John Wiley; 1990.

Breslow, N. Extra-Poisson variation in log-linear models. *Applied Statistics*. 33:38–44; 1984.

Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52(3):345–370; 1987.

Cameron, A. C.; Trivedi, P. K. Econometric models based on count data: Comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1:29–53; 1986.

Cameron, A. C.; Trivedi, P. K. Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics 46:347–364; 1990.

Carroll, R. J.; Ruppert, D. Transformation and weighting in regression. New York: Chapman and Hall; 1988.

Cox, D.R. Some remarks on overdispersion. Biometrika. 70(1):269–274; 1983.

Cox, D. R.; Lewis, P. A. W. The statistical analysis of series of events. London: Chapman and Hall; 1966.

Cramer, J. S. Econometric applications of maximum likelihood methods. New York: Cambridge University Press; 1986.

Dean, C.; Lawless, J. F. Tests for detecting overdispersion in Poisson regression models. Journal of the American Statistical Association. 84(406):467–472; 1989.

Frome, E. L.; Cragle, D. L.; McLain, R. W. Poisson regression analysis of the mortality among a cohort of World War II nuclear industry workers. Radiation Research 123:138–152; 1990.

Joshua, S. C.; Garber, N. J. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15:41–58; 1990.

Jovanis, P. P.; Chang, H. L. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068:42–51; 1986.

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

Lawless, J. F. Negative binomial and mixed Poisson regression. The Canadian Journal of Statistics. 15(3):209–225; 1987.

Maycock, G.; Hall, R. D. Accidents at 4-am roundabouts. Report 1120. Crowthorne, U.K.: Transport and Road Research Laboratory; 1984.

McCullagh, P.; Nelder, J. A. Generalized linear models. London: Chapman and Hall; 1983.

Miaou, S.-P.; Lum, H. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25:689–709; 1993a.

Miaou, S.-P.; Lum, H. A statistical evaluation of the effects of highway geometric design on truck accident involvements. Transportation Research Record; 1407: 11–23; 1993b.

Miaou, S.-P.; Hu, P. S.; Wright, T.; Davis, S. C.; Rathi, A. K. Development of relationships between truck accidents and highway geometric design: Phase I. Technical Memorandum. Oak Ridge, TN: Oak Ridge National Laboratory; 1991.

Miaou, S.-P.; Hu, P. S.; Wright, T.; Rathi, A. K.; Davis, S.C. Relationships between truck accidents and highway geometric design: A Poisson regression approach. Transportation Research Record 1376:10–18; 1992.

Miaou, S.-P.; Hu, P. S.; Wright, T.; Davis, S. C.; Rathi, A. K. Development of relationships between truck accidents and geometric design: Phase I. FHWA-RD-91-124. Washington, DC: Federal Highway Administration; 1993.

Okamoto, H.; Koshi, M. A method to cope with the random errors of observed accident rates in regression analysis. Accid. Anal. Prev. 21:317–332; 1989.

Roy Jorgensen Associates, Inc. Cost and safety effectiveness of highway design elements. Report 197. Washington, DC: National Cooperative Highway Research Program; 1978.

Saccomanno, F. F.; Buyco, C. Generalized loglinear models of truck accident rates. Transportation Research Record 1172:23–31; 1988.

Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 54:439–447; 1974.

Zegeer, C. V.; Hummer, J.; Reinfurt, D.; Herf, L.; Hunter, W. Safety effects of cross-section design for two-lane road. Volumes I and II. Chapel Hill: University of North Carolina; 1987.

Zegeer, C. V.; Stewart, R.; Reinfurt, D.; Council, F.; Neuman, T.; Hamilton, E.; Miller, T.; Hunter, W. Cost effective geometric improvements for safety upgrading of horizontal curves. Chapel Hill: University of North Carolina; 1990.