Accid. Anal. and Prev. Vol. 25, No. 6, pp. 689–709, 1993
Printed in the U.S.A.

0001-4575/93 $6.00 + .00

MODELING VEHICLE ACCIDENTS AND HIGHWAY GEOMETRIC DESIGN RELATIONSHIPS

SHAW-PIN MIAOU[1] and HARRY LUM[2]

[1]Center for Transportation Analysis, Energy Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6366, Building 5500A, Oak Ridge, TN 37831, USA; [2]Office of Safety and Traffic Operations R&D, Federal Highway Administration, U.S. Department of Transportation, Turner-Fairbank Highway Research Center, 6300 Georgetown Pike, McLean, VA 22101-2296, USA

(*Received* 24 *June* 1992; *in revised form* 10 *February* 1993)

Abstract—The statistical properties of four regression models—two conventional linear regression models and two Poisson regression models—are investigated in terms of their ability to model vehicle accidents and highway geometric design relationships. Potential limitations of these models pertaining to their underlying distributional assumptions, estimation procedures, functional form of accident rate, and sensitivity to short road sections, are identified. Important issues, such as the treatment of vehicle exposure and traffic conditions, and data uncertainties due to sampling and nonsampling errors, are also discussed. Roadway and truck accident data from the Highway Safety Information System (HSIS), a highway safety data base administered by the Federal Highway Administration (FHWA), have been employed to illustrate the use and the limitations of these models. It is demonstrated that the conventional linear regression models lack the distributional property to describe adequately random, discrete, nonnegative, and typically sporadic vehicle accident events on the road. As a result, these models are not appropriate to make probabilistic statements about vehicle accidents, and the test statistics derived from these models are questionable. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships. However, if the vehicle accident data are found to be significantly overdispersed relative to its mean, then using the Poisson regression models may overstate or understate the likelihood of vehicle accidents on the road. More general probability distributions may have to be considered.

1. INTRODUCTION

The empirical relationships between vehicle accidents and highway geometric design variables, such as horizontal curvature, vertical grade, lane width, and shoulder width have been investigated through statistical models in numerous studies (e.g. Roy Jorgensen Associates, Inc., 1978; Zegeer et al. 1987; Okamoto and Koshi 1989; Miaou et al., 1991). The safety questions that these statistical models have been used to address include:

1. Given a section of highway with specified attributes, what vehicle accident involvement rate and accident probability* can be reasonably expected?
2. Given a set of highway geometric design variables, which variables are relatively more critical to the safety performance of the road?
3. What percentage of reduction in vehicle accidents can be expected from various improvements in highway geometric design?

These models have also been used to develop procedures to assess the costs and benefits of potential highway geometric design alternatives (e.g., Zegeer et al. 1990).

Many models have been developed to establish the relationships for different roadway classes, vehicle configurations, and accident severity types. However, most of these models were developed using the conventional linear regression and have been reported to have several unsatisfactory statistical properties in describing vehicle accident events on the road (Jovanis and Chang, 1986; Saccomanno and Buyco 1988). These unsatisfactory properties of the linear regression models have led to the investigation of the Poisson and negative binomial regression models in recent studies (Joshua and Garber, 1990;

*In this study, accident probability refers to the probability of observing "y_i" vehicles involved in accidents on a particular road section i during a period of time, where $y_i = 0, 1, 2, 3, \ldots$ and $i = 1, 2, 3, \ldots, n$.

Miaou et al. 1991). These models are, however, not without limitations.

These developed models were different in (i) the distributional assumption of the occurrences of vehicle accidents; (ii) the choice of geometric design and confounding variables (covariates or explanatory variables); (iii) the choice of the functional form that links the accident rate of a road section to its associated attributes, which can be additive or multiplicative; (iv) the way that road sections are delineated, especially length of road sections; (v) the treatment of vehicle exposure and traffic conditions; and (vi) the consideration of data uncertainties, e.g. the uncertainties of accident location and vehicle exposure data, due to sampling and nonsampling errors.

These regression models have had mixed results, and to date no specific relationships have been widely accepted by the highway engineering and safety community (Transportation Research Board 1987). It is generally agreed, however, that vehicle accidents are complex events involving the interactions of many factors. These factors include not only the road, but also the vehicles, the drivers (or human factors), the traffic (e.g. vehicle speed and congestion level), and the environment (e.g. weather and lighting conditions). To a great extent, research on this important problem has suffered from both a lack of high quality and detailed data and the use of inappropriate statistical techniques to analyze the data.

The objective of this study was to investigate the statistical properties of four regression models—two conventional linear regression models and two Poisson regression models—that have typically been used in previous studies to develop relationships between vehicle accidents and highway geometric design. These models are examined in terms of their underlying distributional assumptions, estimation procedures, functional form, and sensitivity to short road sections. Important issues pertaining to these models, such as the treatment of vehicle exposure and traffic conditions, and potential data uncertainties due to sampling and nonsampling errors, are also discussed. Data from the Highway Safety Information System (HSIS), a highway safety data base administered by the Federal Highway Administration (FHWA), have been employed to illustrate the use and the potential limitations of these models in developing such relationships.

First, four statistical models that have been used in previous studies to establish the relationships are reviewed. Second, properties and possible variations of these models are discussed. Third, data from the HSIS are used to illustrate the use and the limitations of these models. The last section concludes the study.

2. STATISTICAL MODELS

The models described in this section can be applied to any roadway class, vehicle configuration, and accident severity type. For ease of exposition, the following presentation focuses on accidents of all severity types involving all types of vehicles on a particular roadway class.

Consider a set of n road sections of a particular roadway type, say, a rural interstate. Let Y_i be a random variable representing the number of vehicles involved in accidents on road section i during a period of one year, where $i = 1, 2, \ldots, n$.* Further, the actual observation of Y_i during the period is denoted as y_i, where $y_i = 0, 1, 2, 3, \ldots$ and $i = 1, 2, \ldots, n$. The amount of vehicle travel (or vehicle exposure) during the sample year on this road section, denoted by v_i, is usually computed as $365 \times \text{AADT}_i \times l_i$, where AADT_i is the average annual daily traffic (in number of vehicles), and l_i is the length (in miles or kilometers) 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 $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ik})$. 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. For example, to assess the effects of terrain on vehicle accidents, a dummy variable is set equal to 0 if a road section is located in level terrain and set equal to 1 if located in rolling or mountainous terrain.

Four types of models are examined in this paper: an additive linear regression model (R1 model), a multiplicative linear regression model (R2 model), and two multiplicative Poisson regression models (P1 model and P2 model). The general forms of these models and brief descriptions of their estimation procedures are presented in this section.

All models are formulated under the assumption that (i) vehicle miles data and other covariates are free from errors, and (ii) the occurrences of vehicle accidents on different road sections are independent. For comparison purposes, these four models

*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.

†If only accidents involving trucks are of interest, then a typical exposure measure is truck miles, computed as $365 \times \text{AADT}_i \times (\text{T}\%_i/100) \times l_i$, where $\text{T}\%_i$ is the average percentage of trucks in the traffic stream (or percent trucks) on road section i, e.g., 15, and $\text{AADT}_i \times (\text{T}\%_i/100)$ is the "truck AADT" of road section i during the observed year.

Table 1. The four statistical models examined

Model	Formulation	Equation
R1	$Y_i \sim ind\ N(\mu_i, v_i^2\sigma^2)$ where the mean $\mu_i = E(Y_i) = v_i[\mathbf{x}_i'\beta]$ $i=1,2,3,\ldots,n.$	(3)
R2	$\log(Y_i+\delta) \sim ind\ N(\theta_i, \sigma^2)$ where $\theta_i = E(\log(Y_i+\delta)) = \log(v_i) + \beta_1^* + \sum_{j=2}^{k} \beta_j \log(1+x_{ij})$ $i=1,2,3,\ldots,n.$ and δ is a selected small constant (e.g., 0.01, 0.0001).	(6)
P1	$Y_i \sim ind$ Poisson(μ_i) or $p(Y_i=y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$ $i=1,2,\ldots,n.$ where $\mu_i = E(Y_i) = v_i[e^{\mathbf{x}_i'\beta}] = v_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j}]$ $i=1,2,3,\ldots,n.$	(10)&(11)
P2	$Y_i \sim ind$ Poisson(μ_i) or $p(Y_i=y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$ $i=1,2,\ldots,n.$ where $\mu_i = E(Y_i) = v_i\left[\beta_1\left(\prod_{j=2}^{k}(1+x_{ij})^{\beta_j}\right)\right]$ $i=1,2,3,\ldots,n.$	(12)&(13)

are shown collectively in Table 1. Based on these models, the underlying distribution of Y_i and its mean, variance, and coefficient of skewness are shown in Table 2.

R1 Model: An additive linear regression model

$$\frac{Y_i}{v_i} = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i$$

$$\varepsilon_i \sim ind\ N(0, \sigma^2) \quad i = 1, 2, \ldots, n \quad (1)$$

or, equivalently,

$$Y_i = v_i[\mathbf{x}_i'\boldsymbol{\beta}] + v_i\varepsilon_i = v_i\left[\sum_{j=1}^{k} x_{ij}\beta_j\right] + v_i\varepsilon_i$$

$$\varepsilon_i \sim ind\ N(0, \sigma^2) \quad i = 1, 2, 3, \ldots, n \quad (2)$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown regression coefficients, the transpose of which is denoted by $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_k)$, ε_i is an error term, and $\varepsilon_i \sim ind\ N(0, \sigma^2)$ reads as the ε_i are independently and normally distributed with zero mean and constant variance σ^2 (> 0). The normality assumption was used primarily to obtain tests and confidence

Table 2. Underlying distributions of the number of vehicles involved in accidents, Y_i, for the studied models and their mean, variance, and coefficient of skewness

Model	Equations	Distribution of Y_i	$E(Y_i)$ $(=\mu_i)$	$Var(Y_i)$	$Skew(Y_i)$
R1	(1),(2),(3)	Normal	$v_i\left[\mathbf{x}_i'\boldsymbol{\beta}\right]$	$v_i^2\sigma^2$	0
R2	(4),(5),(6)	Log-Normal	$-\delta + v_i\left[\beta_1\left(\prod_{j=2}^{k}(1+x_{ij})^{\beta_j}\right)\right]e^{\frac{1}{2}\sigma^2}$	$(\mu_i+\delta)^2(e^{\sigma^2}-1)$	$\left(e^{\sigma^2}-1\right)^{1/2}\left(e^{\sigma^2}+2\right)$ (>0)
P1	(10),(11)	Poisson	$v_i\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\right]$	μ_i	$\mu_i^{-1/2}$ (>0)
P2	(12),(13)	Poisson	$v_i\left[\beta_1\left(\prod_{j=2}^{k}(1+x_{ij})^{\beta_j}\right)\right]$	μ_i	$\mu_i^{-1/2}$ (>0)

statements about the estimated regression coefficients. It was seldom used to make probabilistic statements about Y_i. Note that whenever appropriate higher order and interaction terms of covariates can be included in eqns (1) and (2) without difficulties.

The model can also be expressed as

$$Y_i \sim ind\ N(\mu_i, \nu_i^2\sigma^2)$$
$$\text{where the mean } \mu_i = E(Y_i) = \nu_i[\mathbf{x}_i'\boldsymbol{\beta}]$$
$$i = 1, 2, 3, \ldots, n. \quad (3)$$

and $E(Y_i)$ reads as the expected value of Y_i. Since $\mathbf{x}_i'\boldsymbol{\beta}$ $(= E(Y_i)/\nu_i)$ represents the expected number of vehicles involved in accidents per vehicle mile (or kilometer) of travel, it is called the "rate function" of the model, and is denoted by λ_i. This model implies that the expected number of vehicles involved in accidents, μ_i, is proportional to vehicle travel, ν_i, and the proportional constant is determined through the rate function.

Using eqn (1), the least squares estimates of the coefficients, which are also the maximum likelihood estimates (MLE) under the normal assumption, can be obtained with most of the standard linear regression computer programs (e.g. Draper and Smith, 1981; Weisberg, 1985). The associated statistics of the estimated coefficients, such as t-statistics, can also be obtained. Early studies using this type of model can be found in the National Cooperative Highway Research Program (NCHRP) Report 197 (Roy Jorgensen Associates, Inc. 1978). The model was also examined in several recent studies, e.g. Okamoto and Koshi (1989), Zegeer et al. (1990), Joshua and Garber (1991), and Mohamedshah, Paniati, and Hobeika (1992).

R2 Model: A multiplicative linear regression model

$$Y_i + \delta = \nu_i\left[\beta_1\left(\prod_{j=2}^{k}(1 + x_{ij})^{\beta_j}\right)\right]e^{\varepsilon_i}$$
$$\varepsilon_i \sim ind\ N(0, \sigma^2) \quad i = 1, 2, 3, \ldots, n. \quad (4)$$

or

$$\log\left(\frac{Y_i + \delta}{\nu_i}\right) = \beta_1^* + \sum_{j=2}^{k}\beta_j\log(1 + x_{ij}) + \varepsilon_i$$
$$\varepsilon_i \sim ind\ N(0, \sigma^2) \quad i = 1, 2, 3, \ldots, n. \quad (5)$$

where δ is a selected small constant (e.g., 0.01, 0.0001), which is added to Y_i to avoid log(0) problem, and $\beta_1^* = \log(\beta_1)$. This model requires that $x_{ij} \geq 0$ for all i and j. One has been added to each covariate (except x_{i1}) to avoid log(0) problem in eqn (5). Without the transformation of adding one to the covariates, the right hand side of eqn (4) will be rendered zero as long as one covariate has value zero, regardless of what the values of other covariates are. Basically, this transformation shifts the "origin" of the covariates from zero to one. Since $(1 + 0)^{\beta_j} = 1$ and $\beta_j\log(1 + 0) = 0$, these covariates do not contribute to the occurrences of accidents at this new origin in eqn (4) and (5). This is a desirable property. For example, horizontal curvature should not be a contributing factor to the occurrences of accidents on tangent road sections. Note that this transformation can be applied to 0, 1 dummy variables without problem. Other transformations are, of course, possible.

The model can also be reexpressed as

$$\log(Y_i + \delta) \sim ind\ N(\theta_i, \sigma^2)$$
$$\text{where } \theta_i = E(\log(Y_i + \delta)) = \log(\nu_i)$$
$$+ \beta_1^* + \sum_{j=2}^{k}\beta_j\log(1 + x_{ij}) \quad i = 1, 2, 3, \ldots, n. \quad (6)$$

As in the R1 Model, the lognormal distributional assumption was used primarily to obtain tests and confidence statements about the estimated regression coefficients. It was not intended to make probabilistic statements about Y_i.

It is important to point out that the expected value of Y_i under the model is:

$$\mu_i = E(Y_i)$$
$$= -\delta + \nu_i\left[\beta_1\left(\prod_{j=2}^{k}(1 + x_{ij})^{\beta_j}\right)\right]e^{\frac{1}{2}\sigma^2} \quad (7)$$

not

$$-\delta + \nu_i\left[\beta_1\left(\prod_{j=2}^{k}(1 + x_{ij})^{\beta_j}\right)\right]. \quad (8)$$

Equation (7) includes an adjustment factor, $\exp(\sigma^2/2)$, which is multiplicative and grows exponentially with the variance σ^2. The estimator in eqn (8), without the adjustment factor $\exp(\sigma^2/2)$, can be shown to provide an estimator for the median, rather than the mean, of the probability distribution of Y_i, and this estimator systematically underestimates the mean of Y_i [e.g. Miller, 1984]. In a typical vehicle accident study, the magnitude of this adjustment factor is quite large because of large σ^2. Therefore,

ignoring this adjustment factor would seriously understate the expected number of vehicle accident involvements. The illustrations in Section 4 show that without this adjustment factor the underestimations are over 80% when the expected total accident involvements across road sections ($\Sigma_i \hat{\mu}_i$) are compared with the observed total ($\Sigma_i y_i$).

Using the linearized model in eqn (5), the least squares estimates of the coefficients, denoted by $\hat{\beta}_1^*$ and $\hat{\beta}_j$, $j = 2, 3, \ldots, k$, estimated variance of the residuals, denoted by $\hat{\sigma}^2$, as well as the estimated variance and t-statistics of the estimated coefficients, can be obtained with a standard linear regression computer program. Substituting the coefficients and residual variance σ^2 in eqn (7) with the least squares estimates gives an estimate of μ_i. That is,

$$\hat{\mu}_i = -\delta + \nu_i \left[\hat{\beta}_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\hat{\beta}_j} \right) \right] e^{\frac{1}{2}\hat{\sigma}^2} \quad (9)$$

where $\hat{\beta}_1 = \exp(\hat{\beta}_1^*)$. This estimate is somewhat biased because $\exp(\hat{\beta}_1^*)$ is a biased estimate of β_1. However, the bias of this estimate is much smaller than that using eqn (8) as an estimator (Miller 1984).

The rate function, λ_i, of this model is the expected number of vehicle involvements, μ_i, in eqn (7) divided by vehicle travel ν_i, which is somewhat complicated when compared to that of the R1 Model because of the small constant δ and the exponential adjustment factor. This model again implies that μ_i is linearly related to vehicle travel ν_i.

These types of models have been used in several recent studies, e.g. Zegeer et al. (1987), Zegeer et al. (1990), and Mohamedshah et al. (1992). However, to our knowledge, most of these studies did not consider the adjustment factor when using the model to estimate or predict accidents. Also, the choice of δ was arbitrary, e.g. $\delta = 0.01$ in Zegeer et al. (1990). In theory, δ can be estimated together with other coefficients in the model. Even so, the interpretation of the estimated δ would be difficult.

P1 Model: A multiplicative Poisson regression model

$$Y_i \sim ind \text{ Poisson}(\mu_i) \text{ or } p(Y_i = y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad i = 1, 2, 3, \ldots, n. \quad (10)$$

where

$$\mu_i = E(Y_i) = \nu_i [e^{\mathbf{x}_i'\boldsymbol{\beta}}] = \nu_i [e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \quad i = 1, 2, 3, \ldots, n. \quad (11)$$

This model assumes that Y_i, $i = 1, 2, \ldots, n$, are independently and Poisson distributed with mean μ_i. As in the R1 and R2 Models, the expected number of vehicles involved in accidents μ_i in this model is proportional to vehicle travel ν_i. The model, however, assumes an exponential rate function, $\lambda_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, 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 et al. 1990). Note that whenever appropriate higher order and interaction terms of covariates can be included in eqn (11) without difficulties.

The Poisson distributional assumption is used to obtain tests and confidence statements about the estimated regression coefficients and, unlike the R1 and R2 Models, this distribution can also be used to make reasonable probabilistic statements about Y_i in many cases.

The regression coefficients of this model can be estimated using the maximum likelihood method (Cramer 1986), the quasi-likelihood method (McCullagh and Nelder 1983), or the generalized least squares method (Carroll and Ruppert 1989). For the asymptotic variance and t-statistics of the estimated coefficients, which are based on the second derivative of the loglikelihood function, see Cramer (1986) for reference.

This model has been used to develop truck accidents and highway geometric design relationships in Miaou et al. (1991, 1992). It has also been used in other areas of highway safety studies. For example, Jovanis and Chang (1986) used the model to examine the relationship between vehicle accidents and vehicle miles of travel; Saccomanno and Buyco (1988) applied the model to relate vehicle accident rates with different traffic volumes, truck types, hour of day, and driver ages.

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. 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 that the Poisson model indicated.

In vehicle accidents-geometric design studies, the overdispersion could come from several possible sources. Some sources were identified in Miaou et al. (1991):

1. Omitted variables—Ideally, all of the relevant variables should be considered when developing relationships between vehicle accidents and highway geometric design. In practice, some of these factors may not be available for individual road sections, especially, detailed vehicle exposure data by vehicle type, time of day, and weather. Also, many vehicle accidents are directly or indirectly related to human factors, and these factors are not likely to be found in any data base for individual road sections. This means that some variables that may have influences on the occurrences of accidents will not be included in the model, especially, those qualitative types of variables.
2. Uncertainties in vehicle exposure data and traffic variables—AADT and percent trucks are often estimated using data collected by a highway sampling system, and therefore, the data are subject to sampling errors (e.g. daily, day of week, seasonal and spatial variations) as well as nonsampling errors (e.g. vehicle axle counting and classification errors).
3. Nonhomogeneous roadway environment—Roadway environment (including lighting and weather conditions) and traffic conditions may not be homogeneous on each road section during a sample period. For example, vehicle accident involvement rate during daytime and nighttime may be different, failing to disaggregate daytime and nighttime vehicle accidents in the analysis may introduce extra variations.

The consequences of ignoring the extra variations in the Poisson regression models are that consistent estimates, such as the MLE, of the regression coefficients under the Poisson model, are still consistent; however, the variances of the estimated coefficients 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 coefficient $\boldsymbol{\beta}$, but we may overstate the significance levels of the estimated coefficients (Cameron and Trivedi, 1990). (Note that this is assuming that the rate function in eqn (11) is correct.) Following Wedderburn (1974), to correct for the overdispersion problem for the P1 model, one can assume that the variance of Y_i is $\tau\mu_i$ instead of μ_i as that originally assumed in the Poisson model, where τ is called the overdispersion parameter. Furthermore, the overdispersion parameter τ can 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 in our case), and k is the number of unknown regression coefficients in the Poisson regression model. The Pearson's X^2 statistic is computed as $\Sigma_i(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$. A better estimate of the asymptotic t-statistic for each regression coefficient is $\tau^{-1/2}$ times that obtained from the original Poisson regression model based on the maximum likelihood method (Agresti 1990).

Other ways to overcome this overdispersion problem are available. A simple way is to use the negative binomial distribution. This distribution allows the variance to exceed the mean. The negative binomial regression models were used in Miaou et al. (1991) to establish truck accidents-geometric design relationships. Although the negative binomial regression model is more general than the Poisson model, it requires more extensive computations to estimate model coefficients and to generate inferential statistics. Furthermore, the statistical properties of different estimators, e.g. the MLE and moment estimators, of the negative binomial regression model under different sample sizes have not yet been fully investigated (Lawless 1987). Currently, many statistical studies are under way attempting to modify existing distribution functions within the exponential family to allow more flexible mean-variance relationships (e.g. Efron 1986; Gelfand and Dalal 1990).

P2 Model: A multiplicative Poisson regression model

$$Y_i \sim ind\ \text{Poisson}(\mu_i) \text{ or } p(Y_i = y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad i = 1, 2, 3, \ldots, n. \quad (12)$$

where

$$\mu_i = E(Y_i) = \nu_i \left[\beta_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\beta_j} \right) \right] \quad i = 1, 2, 3, \ldots, n. \quad (13)$$

This model is different from the P1 Model only in the rate function. Instead of using an exponential rate function as in the P1 Model, this model uses a multiplicative function similar to that used in the R2 Model. Therefore, as in the R2 Model, this model requires that $x_{ij} \geq 0$ for all i and j. Most discussions under the P1 Model apply to the P2 Model as well. This type of model has been

examined in a recent study by Joshua and Garber (1990).

3. PROPERTIES AND VARIATIONS OF THE MODELS

Several unsatisfactory properties pertaining to the linear regression models (R1 and R2 Models) in analyzing vehicle accident data have been reported (e.g. Jovanis and Chang, 1986; Saccomanno and Buyco 1988; Miaou et al. 1991). These properties relate mainly to the underlying distributional assumption of the linear regression models (i.e. the normal assumption). Although the Poisson regression models (P1 and P2 Models) have been advocated in recent studies, these models are certainly not without limitations. Particularly, the variance of a Poisson model is constrained to be equal to its mean. In this section, we discuss some important observations that have been made about vehicle accident events; then look into the four examined models to see if they have the necessary statistical properties to describe these observations. Further, we discuss some possible variations and practical considerations of these models.

Distributional assumption

The occurrences of vehicle accidents are discrete random events. That is, the number of vehicles involved in accidents on a given road section during a period of time is probabilistic in nature and is a nonnegative integer value. Furthermore, the occurrences of vehicle accidents are typically sporadic across the road network. In most vehicle accidents-geometric design studies, especially truck-related studies, the analyst is 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 10,900 road sections they studied had no reported vehicle accidents in a five-year period; in another study by Miaou et al. (1991), over 70% of the road sections have no reported accidents involving trucks during a one-year period. This suggests that for a period of several years most of the road sections considered, especially those in rural areas, would have a higher probability of being observed with no accident than with, say, more than one accident. In other words, the underlying distribution of the occurrences of vehicle accidents on most of the road sections is positively (or rightly) skewed. The normal distribution, which is a symmetrical distribution, is a poor approximation under this condition.

With above observations in mind, the statistical properties of the four examined models are discussed as follows.

1. *Discreteness and Nonnegativity*
   The use of a continuous distribution, such as the normal distribution, is at best an approximation to a truly discrete process. In addition, the use of the normal distribution permits the prediction of a negative number of accidents. The Poisson distribution, on the other hand, is a natural initial candidate distribution for such random discrete and, typically, sporadic events. However, if the accident data are found to be significantly overdispersed relative to a Poisson model, then using the Poisson model to make probabilistic statements about vehicle accidents may overestimate or underestimate the likelihood of vehicle accidents on the road. More general probability distributions may have to be considered.
2. *Mean-Variance-Skewness Relationships* (Table 2)
   Mean, $E(Y_i)$—One criticism that has long been made about the R1 Model is that there is no assurance that the estimate of $E(Y_i)$ will always be nonnegative. In fact, many models of this type presented in the literature can be shown to estimate negative numbers of accidents under some conditions. An example was given by Saccomanno and Buyco (1988). This weakness is very undesirable in practice. The estimates of $E(Y_i)$ from the R2 Model is likely to be positive as long as δ is small and $x_{ij} \geq 0$ for all i and j. The P1 Model ensures that $E(Y_i)$ will always be nonnegative; while the P2 Model will also ensure nonnegativity as long as $x_{ij} \geq 0$ for all i and j.

   Variance, $Var(Y_i)$—All four models assume that $Var(Y_i)$ varies from road section to road section. The variances, however, relate to vehicle exposure ν_i and the rate function λ_i of the model in different ways (see Table 2). In the R1 Model, $Var(Y_i)$ is proportional to ν_i^2, and is unrelated to its rate function. In the R2 Model, $Var(Y_i)$ relates to μ_i^2, while in the P1 and P2 Models, it relates to μ_i. Therefore, $Var(Y_i)$ of the R2, P1, and P2 Models depend on their rate functions and, thus, involves unknown regression coefficients. (Recall that $\mu_i = \lambda_i \nu_i$ and λ_i is a function of unknown regression coefficients). All

four models assume that $Var(Y_i)$ increases as vehicle exposure ν_i increases. However, in the R1 and R2 Models it grows quadratically with ν_i (i.e., $Var(Y_i) \propto \nu_i^2$), while in the P1 and P2 Models it grows linearly with ν_i. As we will soon discuss, the variance function of each model is closely related to the way that accidents on each road section are weighted in the estimation of model coefficients.

Skewness, $skew(Y_i)$—The R1 Model assumes that Y_i is symmetrically distributed on any road section i ($skew(Y_i) = 0$). The R2 model, on the other hand, assumes a positive skewness coefficient which is constant across all road sections. Both the P1 and P2 Models suppose a positive skewness coefficient which varies from road section to road section, depending on their means ($skew(Y_i) = \mu_i^{-\frac{1}{2}}$). Also, as mean $E(Y_i)$ increases, either as a result of an increase in vehicle exposure ν_i or an increase in the rate function λ_i, the skewness coefficient of the Poisson models decreases. This is a desirable property for two reasons: (i) We expect the normal distribution to approximate the true probability distribution of Y_i well when $E(Y_i)$ is reasonably large (e.g. $E(Y_i) > 10$); and (ii) We expect the probability distribution of Y_i for road sections with very small $E(Y_i)$ (e.g. $E(Y_i) < .1$) to be highly positively skewed. In developing vehicle accident-geometric design relationships, we may have $E(Y_i)$ varying over a wide range across road sections, e.g. from 0.01 to 10. This is especially true when studying the relationship in urban and suburban areas. Therefore, we need a distribution function that can provide a wide range of skewness coefficient value. When compared with the P1 and P2 Models, both the R1 and R2 Models lack the distributional flexibility to adequately model vehicle accidents in terms of their skewness coefficient.

Because of the unsatisfactory properties described above about the R1 and R2 Models, these two models can rarely be used to make probabilistic statements about Y_i. Also, the test statistics derived from these two models, which must rely on the normal (or lognormal) assumption, are questionable.

Total number of expected vs. observed vehicle accident involvements ($\Sigma_i \hat{\mu}_i$ vs $\Sigma_i y_i$)

As shown in Appendix A, based on the maximum likelihood estimation method the expected total number of accident involvements from the P1 and P2 Models is guaranteed to be equal to the observed total (except for some small numerical errors). The R1 and R2 Models, on the other hand, do not have such a property. The negative binomial regression models, which were not examined in this study, but were used in Miaou (1991), also do not have this property under the maximum likelihood estimation method. This can be seen from eqns (2, 3) in Lawless (1987). Note that the assumption of the R1 Model can be modified to possess this property. For example, instead of assuming that the residuals, ε_i, $i = 1, 2, \ldots, n$, have constant variance σ^2 in eqn (1), one can assume that the variance is $\nu_i^{-1}\sigma^2$. This method was tested by Zegeer et al. (1990, p. 94).

Given that Y_i is a random variable, the total number of vehicles involved in accidents, $\Sigma_i Y_i$, is also a random variable with some mean and variance. Under the P1 Model, $Var(\Sigma_i Y_i) = \tau(\Sigma_i \mu_i)$, and a reasonable estimate of the standard deviation would be $\hat{\tau}^{1/2}(\Sigma_i y_i)^{1/2}$. For the truck accident illustration in the next section, this standard deviation is about 4% of the total number of observed truck accident involvements ($\Sigma_i y_i$). Therefore, in theory, there is no need for us to insist on having a model which constrains the expected total ($\Sigma_i \hat{\mu}_i$) to be equal to the observed total, as in the Poisson model. Perhaps we can accept a model that have $\Sigma_i \hat{\mu}_i$ close to $\Sigma_i y_i$ within, e.g. one standard deviation of $\Sigma_i y_i$ (which is about 4% of $\Sigma_i y_i$ in our illustration in the next section). However, in practice, when aggregated over a large geographical area, we usually found $\Sigma_i y_i$ to be extremely stable over time (after adjusting for weather, long-term trend, etc.). For example, Fridstrom and Ingebrigtsen (1991) offered the following observation:

> Road casualties are random events. Each single accident is unpredictable in the very strong sense that, had it been anticipated, it would most probably not have happened.
>
> Yet the number of accidents recorded with reasonable large geographical units exhibits a striking stability from one year to the next . . . Although the single event is all but impossible to predict, the collection of such events may very well behave in a perfectly predictable way, . . .

This suggests that $(Var(\Sigma_i Y_i))^{1/2}$ should be quite small when compared to $E(\Sigma_i Y_i)$. It is therefore reasonable to check if $\Sigma_i \hat{\mu}_i$ from a model is reasonably close to $\Sigma_i y_i$. If not, further investigations on the adequacy of the developed model are necessary.

Delineation of road sections

How should a stretch of roadway be delineated into road sections for studying vehicle accidents and geometric design relationships, based on both high-

way design and statistical considerations? Several methods have been considered, e.g. fixed-length sections, say, every 0.1 mi or 1.0 mi (0.16 km or 1.6 km), or homogeneous sections, defined as sections that are homogeneous in major geometric design and traffic characteristics (Okamoto and Koshi 1989; Miaou et al. 1991).

This question has been raised and investigated in several studies in a linear regression context (using the R1 Model) (e.g. Okamoto and Koshi 1989; Zegeer et al. 1990). In general, these studies found that short road sections had undesirable impacts on the estimation of their linear regression models. Longer sections (e.g. greater than 1.0 mi or 1.6 km) were, therefore, suggested to be preferred. However, given that most of the curved and graded sections were relatively short, the analysts oftentimes were unable to find a sufficient number of long and homogeneous road sections, which exhibited wide enough variations in geometric design variables, to study the relationships. In addition, the potential bias introduced by arbitrarily eliminating short road sections was not known. To overcome this problem, many analysts have chosen to keep long road sections and not to insist on having homogeneous road sections. Instead, they allowed road sections to be nonhomogeneous in horizontal curvatures and vertical grades. That is, one road section may have contained multiple curves and multiple grades. To characterize the horizontal and vertical alignments of such road sections, surrogate measures were used in many studies. These surrogate measures were devised as some functions of the curvature degrees and grade percentages along the length of each road section. Some example surrogate measures can be found in Joshua and Garber (1990), Miaou et al. (1992), and Mohamedshah et al. (1992). However, these surrogate measures are not easy to interpret in a design context. The main reason is that these surrogate measures are not unique, i.e. different combinations of curves and grades can result in the same values. Therefore, it may be difficult for design engineers to incorporate these measures into their current practice. Furthermore, it is difficult to consider the effect of "length of curve," "length of grade," and continuous geometric design conditions, such as sharp horizontal curves following long segments of straight alignment, on vehicle accident rate in the model when surrogate measures are used.

The variance functions described earlier have provided some clues as to why linear regression models are so sensitive to short road sections. In the R1 and R2 Models, $Var(Y_i)$ grows quadratically with ν_i and, thus, grows quadratically with section length l_i (i.e. $Var(Y_i) \propto l_i^2$). (Recall that $\nu_i = 365 \times \text{AADT}_i \times l_i$). In estimating the coefficients of linear regression models, the square of the deviation of each observation from its expected value (or the square of the residual) is weighted by the reciprocal of its variance. Therefore, in estimating the coefficients, for road sections with the same AADT the linear regression models put considerably more weights on accidents observed on short road sections than those observed on long sections, when compared with the Poisson regression models. As indicated earlier, most of the curved and graded sections are relatively short, and we expect the accident involvement rates of these sections to be higher than tangent and level road sections. As a result, when a large number of short road sections is included, the overall estimated accident involvement rate would tend to be inflated in the linear regression models.

To illustrate the difference between the Poisson regression models and the linear regression models in their treatment of short road sections, a hypothetical example is designed and presented in Appendix B. This example considers a set of n homogeneous road sections. One of the homogeneous road sections, which had one observed accident, is selected and then randomly divided into m smaller subsections with various lengths. This gives us a new set of $n + m - 1$ homogeneous road sections. It is demonstrated in this example that in the linear regression models it is not possible to tell whether the estimated coefficients based on the original n road sections will be close to those based on the new $n + m - 1$ road sections. In addition, if the observed accident happened to be located on a very short subsection, the new estimates can blow up (see the Appendix B). When the same design is applied to the Poisson regression models, it is shown that the new MLE will be the same as the original MLE regardless of how this selected homogeneous road section is subdivided.

The hypothetical example above suggests the following:

1. A very short road section can have a detrimental impact on the estimation of model coefficients in the linear regression models, but not in the Poisson regression models.
2. In the set of n road sections considered, if some of the road sections have the same covariate values, whether these road sections are aggregated or not before the estimation of coefficients would not make a difference on the final coefficient estimates when the P1 and P2 Model are used, but could

make a big difference when the R1 and R2 Models are applied.

Treatment of vehicle exposure and traffic conditions

The four models presented in Section 2 are all formulated in such a way that the expected number of vehicles involved in accidents on road section i, i.e. $E(Y_i)$, is proportional to its vehicle exposure, ν_i. Statistically, we have $\log(\mu_i) \propto \beta_0 \log(\nu_i)$ with β_0 being set equal to 1. (Note that in the R2 Model we have $\log(\mu_i + \delta) \propto \beta_0 \log(\nu_i)$). Such a term is sometimes called an "offset" (McCullagh and Nelder 1983). If other covariates, $x_{ij}, j = 1, 2, \ldots, k$, are not correlated with ν_i, one possible way of testing whether ν_i is a good measure of vehicle exposure is to replace ν_i in these models with $\nu_i^{\beta_0}$, where β_0 is an unknown constant to be estimated. Under the condition, if ν_i is indeed a good exposure measure, the estimate of β_0 should be close to unity. For example, by revising the P1 Model we obtain the following new model, say, the P3 Model.

P3 Model: A revised multiplicative Poisson regression model

$$Y_i \sim ind \text{ Poisson}(\mu_i) \text{ or } p(Y_i = y_i) = p(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad i = 1, 2, 3, \ldots, n. \quad (14)$$

where

$$\mu_i = E(Y_i) = \nu_i^{\beta_0}[e^{\mathbf{x}_i\boldsymbol{\beta}}] = \nu_i^{\beta_0}[e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \quad i = 1, 2, 3, \ldots, n. \quad (15)$$

This type of model has been used, e.g. by Maycock and Hall (1984) to study vehicle and pedestrian accidents at roundabouts.

When ν_i is correlated with some of the covariates (or some important omitted variables), then the estimate of β_0 may not be close to unity. For example, for urban highways we may want to include some traffic density measures as covariates in the model that allow us to better determine the "congestion-related" vehicle accidents, and these covariates are likely to be functions of AADT and highway capacity and to some extent will be correlated with the exposure measure—vehicle miles. Nevertheless, we still recommend that β_0 be estimated as part of the model diagnostic checking exercises. Further investigations, such as checking the correlation between vehicle miles and other covariates and checking the accuracy of vehicle miles, are required when $\hat{\beta}_0$ is not reasonably close to 1. For the R1 Model, a nonlinear least squares estimation procedure has to be used for estimating β_0; while for the other three models, no changes in estimation procedure are required.

The basic measure of vehicle exposure is the total amount of time that vehicles travel on the road. The other covariates, on the other hand, are intended to reflect the conditions to which these vehicles are exposed during the time of their travel. By assuming that the average vehicle's speed across road sections is a constant, one is able to use the vehicle miles traveled on road sections as a surrogate measure to represent the relative amount of time that vehicles have traveled on different road sections. This assumption would hold better if the road sections under consideration belong to one roadway class (with urban-rural separation), where the average vehicle's speed on different road sections is expected to be approximately equal. If road sections from different roadway classes, typically with different average vehicle speeds, are used together in the analysis, then vehicle miles alone may not be able to appropriately reflect the differences in vehicle exposure between road sections that belong to different roadway classes. To give an example, consider two road sections, say A and B, which had the same total vehicle miles during an observed period. However, road section A had an average vehicle speed of 60 mph (96 kph), while road section B was 40 mph (64 kph). The total amount of time vehicles traveled on road section B is expected to be about 1.5 times longer than that on road section A. Vehicle miles measure alone does not reflect this kind of difference. Some modifications are, therefore, needed in order to make vehicle-mile a compatible measure of vehicle exposure among different roadway classes.

There is a need to make a distinction between vehicle exposure and traffic conditions in the model. Conceptually, the former measures the "opportunity" of accident involvements, while the latter describes the conditions (e.g. congestion levels, traffic density) to which vehicles are exposed, which is similar to the exposure of vehicles to environmental conditions, such as weather and lighting conditions. Plausible measures of traffic conditions include AADT per lane (Miaou et al. 1992), standard deviation of vehicle speeds, volume to capacity (V/C) ratio (FHWA 1987), and vehicle composition in the traffic stream. For a particular road section, the higher the vehicle density and the greater the difference in vehicle speeds, the greater the chance for a vehicle to be involved in a conflicting position with

other vehicles when negotiating its way through the road section.

The above distinctions are rarely made in vehicle accidents-geometric design studies. In many studies of this kind the vehicle exposure concept was not explicitly utilized in developing the relationships. For example, some studies used vehicle accident involvements per mile of road section as dependent variable (e.g. Zegeer et al. 1987; Zegeer et al. 1990; Mohamedshah et al. 1992). In a study on truck accidents by Joshua and Garber (1990), instead of using truck miles explicitly to represent truck exposure, they treated AADT, section length, and percent trucks as separate covariates in their models.

Sampling and nonsampling errors

One of the well-known potential problems with accident data is underreporting (Hauer and Hakkert 1988). Not all accidents are reported or recorded, especially minor accidents. In practice, one way of dealing with this problem is to assume that the degree of underreporting is common to all road sections and make an appropriate correction outside the model.

Another potential problem is the location of an accident, which is often estimated by the police, and occasionally the accident is roughly assigned to the nearest milepost of the route where it occurred. Therefore, assigning vehicle accidents to very short road sections is more susceptible to locational error than assigning to longer road sections. Also, it is possible that an accident begins at one road section and comes to a stop at the following road section. This uncertainty in accident location creates a so-called "errors in dependent variable" problem. As a result, extra model uncertainties are introduced. No study has been found in this particular area.

Vehicle exposure data, computed from AADT and percent vehicles, in the United States come primarily from FHWA's Highway Performance Monitoring System (HPMS), a highway sampling system statistically designed to obtain physical, traffic, and operational information on national highways from a small portion of selected highway sections (FHWA 1987). Both AADT and percent vehicles are subject to sampling errors (e.g. daily, day-of-week, seasonal, and spatial variations) and nonsampling errors (e.g. vehicle axle counting and vehicle classification errors). This area is rarely studied. Miaou et al. (1991) attempted to include the sampling errors of vehicle exposure data in their negative binomial regression models.

4. ILLUSTRATIONS

Data source

To illustrate the use and the potential limitations of these models, data from the HSIS were employed to develop relationships between *truck* accidents and key highway geometric design variables. The HSIS currently has data from five states. A general description of the HSIS data base is given in Council and Paniati (1990). Specifically, accidents involving large trucks* on rural interstate highways from a midwest state were used. Among the five HSIS States, this state was considered to have the most complete information on highway geometric design (Miaou et al. 1991). 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 illustration, data from 1985 to 1987 were used.

Data were stored in six files: roadlog, horizontal curvature, vertical grade, accident, vehicle, and occupant files. Thus, these files had to be linked before any analysis could be performed. Each record in the roadlog file represented a homogeneous section in terms of its cross-sectional characteristics, such as number of lanes, lane width, shoulder width, median type and width, AADT, and percent trucks. However, these road sections were not necessarily homogeneous in terms of their horizontal curvatures and vertical grades. Road sections in the horizontal curvature and vertical grade files, on the other hand, were delineated in such a way that they were homogeneous in terms of their horizontal curvatures and vertical grades, but not necessarily in terms of other road characteristics.

Therefore, after matching road sections in the horizontal curvature and vertical grade files with the road sections in the roadlog file, each road section in the road inventory file may have contained more than one horizontal curvature or vertical grade. In this illustration, those road sections with multiple curvatures and grades were further disaggregated into smaller subsections such that each subsection contained a unique set of horizontal curvature and vertical grade. Each subsection, which was totally homogeneous in cross-sectional characteristics, horizontal curvature, and vertical grade, was then considered as an independent road section in the model.

*Trucks with gross vehicle weight rating of 10,000 lbs or over.

Accidents, characteristics of road sections, and covariates

The time period considered was one year, which means that the same road section, even if nothing had changed, was considered as three independent sections—one for each year from 1985 to 1987. This allowed the year-to-year changes on highway geometric design and traffic conditions to be considered in the model. There was a total of 4,983 homogeneous road sections during the three-year period, which were considered to have reliable data. These road sections constituted 8,668 lane-miles of roadway. Data for each year contained roughly one-third of the total sections and lane-miles. The section lengths varied from 0.01 to 7.77 mi (0.016 to 12.43 km)—with an average of 0.44 mi (0.70 km). Descriptive statistics of these 4,983 road sections on truck accident involvements and truck miles traveled are given in Table 3.

During the three-year period, 927 large trucks were 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 1,054 million truck miles, the overall truck accident involvement rate was 0.88 truck involvements per million truck miles. These accidents occurred on only 13% of the 4,983 road sections. The maximum number of trucks involved in accidents on an individual road section was 7. On average, each section had 0.19 trucks involved in accidents in one year.

The covariates considered for individual road sections and their definitions are also presented in Table 3. 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, 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 were 12 ft. in lane width, over 90% of them had 4 lanes, and almost all road sections had a paved outside (or right) shoulder width of 10 ft. per direction, we were 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 mph or 16 kph), the accident rate increases (Roy Jorgensen Associates, Inc. 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 (x_{i6}) and length of original vertical grade (x_{i8})—were considered. As indicated earlier, each curve or grade considered in the model may have been subdivided from a longer curve or grade for achieving total homogeneity. Thus, for each road section in the model, these two covariates were defined as the length of the original undivided curve or grade to which this section belonged. In addition, these two covariates were defined only for curves with horizontal curvatures greater than 1 degree and sections with grade greater than 2%. (Note that these two covariates were 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 was based on an assumption that the length of a mild curve or grade has no aggravated effect on truck accident involvements.

Model results

The estimated regression coefficients of the four studied models (R1, R2, P1, and P2 Models) and the associated t-statistics are presented in Table 4. The loglikelihood function evaluated at the estimated coefficients, $L(\hat{\boldsymbol{\beta}})$, and the Akaike Information Criterion (AIC) value (e.g. Bozdogan, 1987) for each model are also given in the table.* Note that $AIC = -2L(\hat{\boldsymbol{\beta}}) + 2k$, where k is the total number of unknown regression coefficients in the model. The Wedderburn's overdispersion parameters for the Poisson regression models are computed to adjust the estimated t-statistics based on the MLE. Furthermore, the expected total number of trucks involved in accidents across road sections is compared with the observed total. To check whether the estimated truck miles could provide a reasonable exposure measure for truck accident involvements in this illustration, the P3 Model was estimated as well.

For the R2 Model in Table 4, δ was selected to be 0.01, as in Zegeer et al. (1990). To see how the selection of δ value might affect the coefficient estimates, the R2 Model was reestimated with different δ values (10^{-6}, 10^{-4}, and 0.02) and the results are presented in Table 5.

In order to illustrate the effect of short road sections on the estimation of different models, we removed road sections with section length less than or equal to 0.05 mi (0.08 km). As a result, a total of 762 road sections were eliminated. The remaining 4,221 road sections, which had 900 truck accident

*Estimated models with high loglikelihood function and low AIC values are preferred.

Table 3. Variable definitions and summary statistics of the 4,983 road sections[1]

Variable	Notation & Definition (for section i)	Min	Max	Mean	% Zeros[2]
Number of Trucks Involved in Accidents	y_i	0	7	0.19	87
Section Length (in mi)	l_i	0.01	7.77	0.44	0
Truck Miles or Truck Exposure (in 10^6 truck-miles)	$v_i = [365 \times AADT_i \times (T\%_i/100) \times \ell_i]/10^6$, where $T\%_i$ is percent trucks (e.g., 15).	6.77×10^{-4}	4.46	0.21	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				
AADT per Lane (in 1000's of vehicles), a surrogate variable to indicate traffic conditions or traffic density.	$x_{i4} = AADT_i/(\text{number of lanes})_i/1000$	0.35	10.58	1.69	0
Horizontal Curvature (in degrees per 100-ft arc)	x_{i5}	0	12.00	0.99	67
Length of Original Horizontal Curve (in mi) from which this curve was subdivided for creating homogeneous sections; only for sections with curvature > 1 degree.	x_{i6} = Length of original curve, if $x_{i5} > 1$ $= 0$, if $x_{i5} \leq 1$	0	0.96	0.05	81
Vertical Grade (in percent)	x_{i7}	0	8.00	2.15	20
Length of Original Vertical Grade (in mi) from which this section was subdivided for creating totally homogeneous sections; only for sections with grade > 2 percent.	x_{i8} = Length of original grade, if $x_{i7} > 2$ $= 0$, if $x_{i7} \leq 2$	0	3.85	0.22	74
Deviation of Inside Paved Shoulder Width from an "ideal" width of 12 ft per direction (in ft).	$x_{i9} = \max\{0, 12 - \text{paved inside shoulder width}\}$	4.00	12.00	8.13	0

[1] All of the sections are 12 ft in lane width; over 90% of the sections have 4 lanes; and all road sections have paved outside shoulder width of 10 ft. Total number of trucks reported to be involved in accidents = 927; Total highway lane-miles = 8,668; Total truck miles traveled = 1,054 million truck miles.

[2] Percentage of road sections that have values of zero.

(1 mi = 1.6 km; 1 ft = 0.3048 m)

Table 4. Estimated coefficients of the studied models and associated statistics. (4,983 rural Interstate road sections; section length ≥ 0.01 miles)

Coefficient	R1 Model	R2 Model	P1 Model	P2 Model	P3 Model
β_0 Truck miles or exposure (10^6)					0.895139 (24.21)
β_1 Dummy intercept	2.317120 (1.81)	0.115623 (1.48)	-0.542106 (-1.10)	0.232342 (0.56)	-0.477960 (-1.06)
β_2 Dummy variable for 1986	-0.675292 (-2.17)	-0.626610 (-7.19)	-0.362042 (-3.52)	-0.520477 (-3.45)	-0.326795 (-3.45)
β_3 Dummy variable for 1987	-0.317802 (-1.01)	-0.568105 (-6.46)	-0.341286 (-3.38)	-0.494596 (-3.34)	-0.298762 (-3.20)
β_4 AADT per lane (10^3)	0.073134 (0.77)	-0.673813 (-10.55)	0.043083 (1.81)	0.184146 (1.76)	0.051066 (2.34)
β_5 Horizontal curvature	0.156019 (2.03)	0.462109 (9.44)	0.179679 (5.38)	0.381715 (3.36)	0.141124 (4.24)
β_6 Length of original curve	1.048150 (0.92)	-0.780708 (-2.54)	0.006385 (0.02)	0.185697 (0.34)	0.058679 (0.17)
β_7 Vertical grade	0.152990 (1.30)	0.394915 (7.20)	0.119672 (2.74)	0.192734 (1.91)	0.110336 (2.80)
β_8 Length of original grade	-0.147874 (-0.39)	0.043079 (0.40)	0.116193 (1.09)	0.471076 (2.93)	0.121032 (1.23)
β_9 Deviation of inside shoulder width from 12 ft per direction	-0.163698 (-1.06)	0.458425 (1.52)	0.027697 (0.46)	0.478567 (0.60)	0.006633 (0.12)
$L(\hat{\beta})$	-17981.2	-9824.7	-2181.3	-2191.6	-2175.8
AIC Value	35980.5	19667.4	4380.7	4401.2	4371.7
Expected vs. Observed Total Truck Accident Involvements	1189.5 927.0	883.7* 927.0	928.2 927.0	927.5 927.0	927.0 927.0

Notes: (1) Values in parentheses are (adjusted) t-statistics of the coefficients above. Wedderburn's overdispersion parameters for the P1 Model, P2 Model, and P3 Model were estimated to be 1.58, 1.63, and 1.33, respectively. The adjusted t-statistics of these three models were, therefore, computed as the unadjusted t-statistics divided by 1.26 ($=1.58^{1/2}$), 1.28 ($=1.63^{1/2}$), and 1.15 ($=1.33^{1/2}$), respectively.

(2) * The total number of estimated truck accident involvements would be 155.8 if the adjustment factor in Eq. (7) is ignored.

(3) In the R2 Model, the small constant δ is set equal to 0.01.

involvements, were used to develop new models. The results are presented in Table 6.

By comparing the estimated coefficients and associated statistics of different models in Tables 4, 5, and 6, the following observations were made:

1. Based on the likelihood function and the AIC values, the Poisson regression models had considerably better performance than the conventional linear regression models. In addition, the P1 Model was favored over

Table 5. Estimated coefficients of the R2 Models with different δ values. (4,983 rural Interstate road sections; section length ≥ 0.01 miles)

Coefficient	$\delta=10^{-6}$	$\delta=10^{-4}$	$\delta=0.01$	$\delta=0.02$
β_1 Dummy intercept	0.000076 (0.58)	0.002966 (0.85)	0.115623 (1.48)	0.201028 (1.62)
β_2 Dummy variable for 1986	-0.682058 (-3.06)	-0.654306 (-4.34)	-0.626610 (-7.19)	-0.622488 (-7.82)
β_3 Dummy variable for 1987	-0.394102 (-1.75)	-0.481182 (-3.16)	-0.568105 (-6.46)	-0.581051 (-7.23)
β_4 AADT per lane (10^3)	0.410656 (2.51)	-0.132009 (-1.19)	-0.673813 (-10.55)	-0.754622 (-12.94)
β_5 Horizontal curvature	0.079347 (0.63)	0.270887 (3.20)	0.462109 (9.44)	0.490617 (10.98)
β_6 Length of original curve	0.441127 (0.56)	-0.170337 (-0.32)	-0.780708 (-2.54)	-0.871640 (-3.11)
β_7 Vertical grade	0.477085 (3.40)	0.435956 (4.59)	0.394915 (7.20)	0.388812 (7.76)
β_8 Length of original grade	0.222984 (0.81)	0.133006 (0.71)	0.043079 (0.40)	0.029589 (0.30)
β_9 Deviation of inside shoulder width from 12 ft per direction	-0.307682 (-0.40)	0.075771 (0.15)	0.458425 (1.52)	0.515334 (1.87)
$L(\hat{\beta})$	-14504.7	-12559.9	-9824.7	-9373.0
AIC Value	29027.5	25137.7	19667.4	18764.0
Expected vs. Observed Total Truck Accident Involvements	1646.8 927.0	373.4 927.0	883.7 927.0	1220.9 927.0

Notes: Values in parentheses are t-statistics of the coefficients above.

the P2 Model. Computationally, we also found that the P1 Model required much less effort than the P2 Model.

2. By comparing the expected and observed total truck accident involvements, several properties pertaining to these models were clear:
   (i) R1 Model—The expected total was considerably greater than the observed total (1,189.5 vs. 927) when short road sections were included (see Table 4). After removing road sections with length less than or equal to 0.05 mi, the new expected total was found to be less than, but much closer to, the observed total (874.9 vs. 900) (see Table 6). This confirms our earlier observation that the estimated overall accident involvement rate was indeed inflated in the R1 Model when short road sections were included.
   (ii) R2 Model—In Table 5, depending on the choice of δ value, the expected totals could be considerably less or higher than the observed totals (e.g. for $\delta = 0.01$, $\Sigma_i\hat{\mu}_i = 883.7$, while for $\delta = 0.02$, $\Sigma_i\hat{\mu}_i = 1{,}220.9$). When the adjustment factor in eqn (7) was ignored, very serious underestimation occurred as expected (e.g. for $\delta = 0.01$, the expected total was only 155.8 in Table 4).

Table 6. Estimated coefficients of the studied models and associated statistics. (4,221 rural Interstate road sections; section length > 0.05 miles)

Coefficient	R1 Model	R2 Model	P1 Model	P2 Model	P3 Model
β_0 Truck miles or exposure (10^6)					0.938188 (25.49)
β_1 Dummy intercept	0.615063 (0.79)	0.065538 (1.40)	-0.709604 (-1.62)	0.154706 (0.55)	-0.672178 (-1.57)
β_2 Dummy variable for 1986	-0.491198 (-2.62)	-0.633460 (-6.92)	-0.348600 (-3.78)	-0.505972 (-3.71)	-0.329828 (-3.62)
β_3 Dummy variable for 1987	-0.464936 (-2.45)	-0.594442 (-6.43)	-0.344338 (-3.80)	-0.501521 (-3.74)	-0.321019 (-3.56)
β_4 AADT per lane (10^3)	0.079664 (1.41)	-0.487529 (-7.25)	0.046935 (2.22)	0.199497 (2.08)	0.051595 (2.47)
β_5 Horizontal curvature	0.206240 (4.22)	0.438168 (8.08)	0.182222 (5.98)	0.374599 (3.56)	0.160212 (4.90)
β_6 Length of original curve	1.235380 (1.78)	-0.629774 (-1.92)	0.019763 (0.05)	0.232429 (0.42)	0.029707 (0.09)
β_7 Vertical grade	0.133424 (1.82)	0.326023 (5.67)	0.120121 (3.06)	0.185148 (2.04)	0.115330 (3.01)
β_8 Length of original grade	-0.035857 (-0.15)	0.242374 (2.15)	0.101627 (1.06)	0.453726 (3.11)	0.100581 (1.06)
β_9 Deviation of inside shoulder width from 12 ft per direction	0.009320 (0.10)	0.532292 (1.67)	0.045014 (0.85)	0.652343 (0.81)	0.033218 (0.63)
$L(\hat{\beta})$	-12753.8	-8180.1	-2058.9	-2069.4	-2057.3
AIC Value	25525.5	16378.2	4135.9	4156.9	4134.5
Expected vs. Observed Total Truck Accident Involvements	874.9 900.0	589.1* 900.0	900.9 900.0	900.6 900.0	900.0 900.0

Notes: (1) Values in parentheses are (adjusted) t-statistics of the coefficients above. Wedderburn's overdispersion parameters for the P1 Model, P2 Model, and P3 Model were estimated to be 1.24, 1.30, and 1.18, respectively. The adjusted t-statistics of these three models were, therefore, computed as the unadjusted t-statistics divided by 1.11 ($=1.24^{1/2}$), 1.14 ($=1.30^{1/2}$), and 1.09 ($=1.18^{1/2}$), respectively.

(2) * The total number of estimated truck accident involvements is 111.2 if the adjustment factor in Eq. (7) is ignored.

(3) In the R2 Model, the small constant δ is set equal to 0.01.

(iii) P1 and P2 Models—In Tables 4 and 6, the expected totals are very close to the observed totals as required by the maximum likelihood estimation procedure of the Poisson regression models.

3. For the R1 Model, in general the estimated regression coefficients in Table 6 have higher t-statistics than those in Table 4. This, again, indicates an inflation in the sum of squares of model residuals due to the inclusion of short road sections. The estimated residual variance, $\hat{\sigma}^2$, for the R1

Model were 80.1 and 49.7, respectively, in Table 4 and Table 6.

When compared to the Poisson regression models, the R1 Model tended to understate the relationships between truck accident involvement rate and geometric design/traffic variables. This was indicated by lower t-statistics in the R1 Model than those in the P1 and P2 Models. The reason was that the R1 Model assumed a larger variance for Y_i than that implied by the P1 and P2 Models. This can be seen from the following example using the results in Table 4: in the R1 Model, $Var(Y_i) = \nu_i^2\sigma^2$, and in the P1 and P2 Models, $Var(Y_i) = \tau\nu_i\lambda_i$. Note that $\hat{\sigma}^2$ for the R1 Model was 80.1 and $\hat{\tau}$ was 1.58 and 1.63 for the P1 and P2 Model, respectively. An average road section in this illustration had $\nu_i = 0.21$ and $\lambda_i = 0.88$; therefore, the estimated variances were 3.53 ($= (0.21)^2 \times 80.1$), 0.29 ($= 1.58 \times 0.21 \times 0.88$), and 0.30 ($= 1.63 \times 0.21 \times 0.88$) for the R1 Model, P1 Model, and P2 Model, respectively.

4. For the P1 and P2 Model, all of the estimated coefficients for the traffic and geometric variables (β_4 through β_9) had the expected sign in both Tables 4 and 6. On the other hand, some of the estimated coefficients in the R1 Model have signs contrary to expectation, e.g. shoulder width and length of original grade in Table 4. In the R2 Model, Table 5 suggests that the choice of δ value can affect the sign of the estimated coefficient, and for $\delta = 0.01$ AADT per lane and length of original curve have signs contrary to what one would usually expect.
5. In the R2 Model, the choice of δ value has significant impact on both the estimation of coefficients and the associated statistics of the estimated coefficients (Table 5).
6. Because the R2 Model and the P2 Model have the same form in the rate function, their coefficient values (except the intercept term) are directly comparable. The examination of these two models in Tables 4 and 6, however, suggested that the estimated coefficients from these two models were very different and some coefficients had different signs.
7. The comparison of the estimated coefficients of the two P1 Models in Table 4 and Table 6 suggested not only that the conclusions reached regarding the significance level of the relationships between truck accidents and the examined traffic and highway geometric variables were consistent, but also that the estimated coefficient values were very close. The same comparison results could be obtained for the P2 Model. This suggests that the Poisson regression models are not sensitive to short road sections as expected.
8. In Tables 4 and 6, the estimated coefficients for the truck miles, $\hat{\beta}_0$, in the P3 Model are 0.895 and 0.938, respectively. While the estimated standard deviations are approximately 0.04 for both coefficients. This sug gests that the estimated coefficients are reasonably close to unity and that the estimated truck miles were a reasonably good exposure measure of truck accident involvements in this illustration.
9. In Table 6, the Wedderburn's overdispersion parameters for the three Poisson regression models are 1.24, 1.30, and 1.18, respectively, which are not far from unity. This suggests that the truck accident data used in this ilustration were moderately overdispersed compared with what the Poisson models have implied. Also, these estimates of overdispersion were smaller than those in Table 4, indicating that the overdispersion measure was somewhat inflated when short road sections were included in the model.
10. Based on the estimated P1 and P2 Models in Table 6, AADT per lane, horizontal curvature, and vertical grade were all positively associated with truck accident involvement rate at a 5% α level. While the increase in the length of curve and the deviation of paved inside shoulder width from 12 ft. were both found to increase the truck accident involvement rate, the coefficients were, however, not as well determined statistically. This indicates that perhaps more data will be needed to better determine the effects of these two variables. The increase in the length of grade was found to increase the truck accident involvement rate under both the P1 and P2 Models. However, the estimated coefficient was found to be significant at a 5% α level under the P2 Model, but not under the P1 Model. This suggests that through the interactions with other covariates the effect of length of grade on truck accident involvement rate becomes more significant.

5. CONCLUSIONS

The statistical properties of four regression models—two conventional linear regression models and two Poisson regression models—were investigated. The potential limitations of these models in developing vehicle accidents and highway geometric design relationships were identified. Roadway and truck accident data from the HSIS were employed to illustrate the use and the limitations of these models to develop such relationships.

It was demonstrated that the conventional linear regression models lack the distributional property to describe adequately random, discrete, nonnegative, and typically sporadic, vehicle accident events on the road. Also, the estimate of vehicle accident involvement rate is sensitive to the length of the road sections considered. Furthermore, there is no assurance that the expected total number of vehicles involved in accidents across road sections from the model would be reasonably close to the observed total. As a result, these models were not appropriate to make probabilistic statements about the occurrences of vehicle accidents on the road, and the test statistics derived from these models were questionable.

The Poisson regression models, on the other hand, possess most of the desirable statistical properties in describing vehicle accident events. One limitation of the Poisson regression model is that the variance of the accident data is constrained to be equal to the mean. As a result, the variances of the estimated model coefficients tend to be underestimated when extra variations or overdispersions exist in the data over a Poisson model. This limitation can be somewhat relaxed by considering the Wedderburn's overdispersion parameter. However, under such overdispersed data, using the Poisson regression models to make probabilistic statements about the occurrences of vehicle accidents may overstate or understate the likelihood of the accidents on the road. More general probability distributions, such as the negative binomial or double Poisson distribution, may have to be considered.

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 authors. The authors would like to thank Patricia Hu, Tommy Wright, and Ed Frome at the Oak Ridge National Laboratory, and the anonymous reviewers of the original manuscript for their helpful suggestions and comments.

REFERENCES

Agresti, A. Categorical Data Analysis. New York: John Wiley; 1990.

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.

Council, F. M.; Paniati, J. F. The Highway Safety Information System. Public Roads 54(3):234–240; December 1990.

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; June 1989.

Draper, N. R.; Smith, H. Applied Regression Analysis. New York: John Wiley; 1981.

Efron, B. Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association 81(395):709–721; Sept. 1986.

FHWA (Federal Highway Administration). Highway Performance Monitoring System Field Manual. Washington, DC: U.S. Department of Transportation; 1987.

Fridstrom, L., and Ingebrigtsen, S. An Aggregate Accident Model Based on Pooled, Regional Time-Series Data. Accident Analysis & Prevention 23(5):363–378; 1991.

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.

Gelfand, A. E.; Dalal, S. R. A note on Overdispersed Exponential Families. Biometrika 71(1):55–64; 1990.

Hauer, E.; Hakkert, A. S. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185:1–10; 1988.

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.

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-arm roundabouts. TRRL Laboratory 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.; 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 prepared by the Oak Ridge National Laboratory. Washington, DC: Federal Highway Administration; March and November 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 Board 71st Annual Meeting, Washington, DC, January 12–16, 1992.

Miller, D. M. Reducing Transformation Bias in Curve Fitting. The American Statistician, 38(2):124–126; May 1984.

Mohamedshah, Y. M.; Paniati, J. F.; Hobeika, A. G. Truck accident models for interstates and two lane rural roads. Transportation Research Board 71st Annual Meeting, Washington, DC, January 12–16, 1992.

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, paper presented at Transportation Research Board 67th Annual Meeting, Washington, DC, January 1988.

Transportation Research Board. Designing safer roads. Special Report 214. Washington, DC: National Research Council; 1987.

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

Weisberg, S. Applied Linear Regression. New York: John Wiley; 1985.

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. Prepared by the University of North Carolina. Washington, DC: Federal Highway Administration and Transportation Research Board; October 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. Prepared by the University of North Carolina. Washington, DC: Federal Highway Administration; May 1990.

APPENDIX A
TOTAL NUMBER OF EXPECTED VERSUS OBSERVED ACCIDENT INVOLVEMENTS

R1 Model: An additive linear regression model

The least squares estimates are obtained by minimizing the residual sum of the squares (RSS) in eqn (1) with respect to (w.r.t.) unknown regression coefficients $\boldsymbol{\beta}$. That is,

$$\min \sum_{i=1}^{n} \left(\frac{y_i}{\nu_i} - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 \quad w.r.t.\ \boldsymbol{\beta}. \tag{A1}$$

To minimize the RSS, the estimated coefficients have to satisfy the following set of "normal equations":

$$\sum_{i=1}^{n} \left(\frac{y_i}{\nu_i} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right) x_{ij} = 0 \quad j = 1, 2, \ldots, k. \tag{A2}$$

Since the first covariate x_{i1} is equal to 1 for all i, the first normal equation requires that $\Sigma_i(y_i/\nu_i) = \Sigma_i(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$. However, there is no assurance that the expected total number of accident involvements $\Sigma_i\hat{\mu}_i$ will be reasonably close to the observed total $\Sigma_i y_i$, where $\hat{\mu}_i = \nu_i(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$.

R2 Model: A multiplicative linear regression model

Using the linearized model in eqn (5), the least squares estimates of the coefficients as well as the estimated variances and t-statistics of the estimated coefficients, can be obtained in a similar way as in the R1 Model. Also, as in the R1 Model, there is no assurance in the estimation process that the expected total number of accident involvements will be close to the observed total.

P1 Model: A multiplicative Poisson regression model

The MLE of the regression coefficients in the P1 Model is obtained by maximizing the following loglikelihood function:

$$L(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{n} p(y_i)\right) = \log\left(\prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}\right) \tag{A3}$$

$$= \sum_{i=1}^{n} \{y_i(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\log(\nu_i) - \nu_i e^{\mathbf{x}_i'\boldsymbol{\beta}} - \log(y_i!)\}$$

The first derivatives of the loglikelihood function w.r.t. the regression coefficients can be shown to be:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \{y_i - \nu_i e^{\mathbf{x}_i'\boldsymbol{\beta}}\} x_{ij} \quad j = 1, 2, \ldots, k, \tag{A4}$$

and must all vanish at the MLE $\hat{\boldsymbol{\beta}}$. Since the first covariate x_{i1} is a dummy variable equal to 1 for all i, the MLE requires that $\Sigma_i y_i = \Sigma_i \nu_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$. That is, the expected total number of accident involvements, $\Sigma_i\hat{\mu}_i$, has to be equal to the observed total $\Sigma_i y_i$, where $\hat{\mu}_i = \nu_i \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$. As indicated in the main text, this is a desirable property in vehicle accident analyses.

P2 Model: A multiplicative Poisson regression model

Similar to the P1 Model, the loglikelihood function can be constructed for the P2 Model, based on which the MLE of the regression coefficients and the asymptotic variances and t-statistics can be obtained. As in the P1 Model, the MLE of the P2 Model ensures that the expected total vehicle accident involvements is equal to the observed total.

APPENDIX B
EFFECTS OF SHORT ROAD SECTIONS

Consider a set of n homogeneous road sections. Here we define a homogeneous road section as a road section having the same covariate values (including AADT and percent trucks) along the length of the section. Suppose that there was one vehicle involved in an accident during the observed period on the last or the nth road section, i.e., $y_n = 1$. Now, let us randomly divide the nth road section into m subsections of various lengths, and the length of each subsection is allowed to be extremely short.

The number of vehicles involved in accidents, section length, vehicle exposure, and the associated covariates of these m subsections are denoted by y_{nq}, l_{nq}, ν_{nq}, and $\mathbf{x}_{nq}$ $(=(x_{n1q}, x_{n2q}, \ldots, x_{nkq})')$, respectively, where $q = 1, 2, \ldots, m$. We have $\Sigma_q y_{nq} = y_n = 1$, $\Sigma_q l_{nq} = l_n$, and $\Sigma_q \nu_{nq} = \nu_n$, and since each is a homogeneous road section, we have $\mathbf{x}_{nq} = \mathbf{x}_n$ and $\text{AADT}_{nq} = \text{AADT}_n$ for all q. Now we have a new set of $n + m - 1$ homogeneous road sections.

For the R1 Model, the new objective function to be minimized for obtaining the least squares estimates is

$$\sum_{i=1}^{n-1} \left(\frac{y_i}{\nu_i} - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 + \sum_{q=1}^{m} \left(\frac{y_{nq}}{\nu_{nq}} - \mathbf{x}_n'\boldsymbol{\beta}\right)^2 \quad \text{(B1)}$$

By differentiating this objective function w.r.t. the regression coefficients and then by setting them equal to zero, we have the following set of "normal equations":

$$\sum_{i=1}^{n-1} \left(\frac{y_i}{\nu_i} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right) x_{ij} + \sum_{q=1}^{m} \left(\frac{y_{nq}}{\nu_{nq}} - \mathbf{x}_n'\hat{\boldsymbol{\beta}}\right) x_{nj} = 0$$
$$j = 1, 2, \ldots, k. \quad \text{(B2)}$$

By comparing this new set of normal equations with the original ones in eqn (A2), one can show that in order for them to obtain the same estimates, the following condition has to be met:

$$\left(\sum_{q=1}^{m} \frac{y_{nq}}{\nu_{nq}}\right) - m(\mathbf{x}_n'\hat{\boldsymbol{\beta}}) = \left(\frac{y_n}{\nu_n}\right) - \mathbf{x}_n'\hat{\boldsymbol{\beta}} = \left(\frac{\sum_{q=1}^{m} y_{nq}}{\sum_{q=1}^{m} \nu_{nq}}\right) - \mathbf{x}_n'\hat{\boldsymbol{\beta}} \quad \text{(B3)}$$

However, in practice this condition is not likely to be satisfied, and there is no assurance that the estimated coefficient will be reasonably close to the original estimates. In addition, if the observed accident happened to be located on a very short subsection, the new estimates can "blow up." (For this particular short subsection, since $l_{nq} \to 0$ and $y_{nq} = 1$, we have $\nu_{nq}(= 365 \times \text{AADT}_n \times l_{nq}) \to 0$ and, therefore, $y_{nq}/\nu_{nq} \to \infty$ in eqn (B2).)

In general, the discussion above on the R1 Model also applies to the R2 Model. However, it is more complicated in the R2 Model because it works with logarithmic scale.

Let us now consider this hypothetical case in the context of the P1 Model. As in eqn (A3), the new loglikelihood function to be maximized is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n-1} \{y_i(\mathbf{x}_i'\boldsymbol{\beta}) + y_i \log(\nu_i) - \nu_i e^{\mathbf{x}_i'\boldsymbol{\beta}} - \log(y_i!)\}$$
$$+ \sum_{q=1}^{m} \{y_{nq}(\mathbf{x}_n'\boldsymbol{\beta}) + y_{nq}\log(\nu_{nq}) - \nu_{nq} e^{\mathbf{x}_n'\boldsymbol{\beta}} - \log(y_{nq}!)\} \quad \text{(B4)}$$

Also, as in eqn (A4), the first derivatives of the loglikelihood function w.r.t. the regression coefficients are

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n-1} \{y_i - \nu_i e^{\mathbf{x}_i'\boldsymbol{\beta}}\} x_{ij}$$
$$+ \sum_{q=1}^{m} \{y_{nq} - \nu_{nq} e^{\mathbf{x}_n'\boldsymbol{\beta}}\} x_{ij} \quad j = 1, 2, \ldots, k, \quad \text{(B5)}$$

and must all vanish at the MLE $\hat{\boldsymbol{\beta}}$. Since $\Sigma_q y_{nq} = y_n$ and $\Sigma_q \nu_{nq} = \nu_n$, eqn (B5) is the same as eqn (A4). Therefore, the new MLE will be the same as the original MLE regardless of how this homogeneous road section is subdivided. This result applies to the P2 Model as well.