Modeling the Relationship of Accidents to Miles Traveled

PAUL P. JOVANIS and HSIN-LI CHANG

ABSTRACT

Consideration of highway safety studies in a time-space domain is used to introduce the concept that different study designs result in different underlying probability distributions describing accident occurrence. Poisson regression is proposed as a superior alternative to conventional linear regression for many safety studies because it requires smaller sample sizes and has other desirable statistical properties. Models are estimated using accident, travel mileage, and environmental data from the Indiana Toll Road. A pooled model including all accidents revealed that accident occurrence increases with automobile vehicle miles of travel (VMT), truck VMT, and hours of snowfall. Segmentation of the data into subsets that describe different types of collisions revealed that automobile accidents are much more sensitive to environmental conditions than are truck accidents. Use of the segmentation technique allowed a much clearer understanding of the effects of travel mileage on accident occurrence than could have been obtained from the pooled data alone.

It is generally recognized that the occurrence of accidents results from the complex interaction of characteristics of the driver, vehicle, roadway, and environment. The number of accidents (accident frequency) is also clearly related to the amount of travel that occurs. Quantity of travel may be measured in any of several ways including hourly volume, average daily traffic, or vehicle miles of travel (VMT), among others. These measures of quantity of travel can be used to describe traffic conditions that exist during exposure to accident risk. A more precise definition of exposure is, "...the amount or opportunity for accidents which the driver or traffic system experiences" (1). This broader interpretation of exposure has led some researchers to explore the effects on accident occurrence of environmental conditions during which the driving occurred (2).

Previous studies relating accident occurrence to level of traffic have used a variety of measures of travel quantity. Belmont (3) found the accident rate (accident per million VMT) for two-lane sections almost linear with hourly traffic flow during daylight. For four-lane divided sections, Leutzbach (4) and Gwynn (5) found that a U-shaped relation exists between accident rate and hourly traffic flow, where the minimum values of the accident rates happened at approximately 600 to 1300 vehicles/hr per two lanes. In another study the accident rate increased rapidly when the traffic volume was below 550 vehicles/hr per two lanes, but showed little variation beyond this flow value (6).

Smeed (7) considered the problem on a much broader scale, studying national yearly accident rates. He found that the total accident rate showed little variation with annual traffic volumes. When he separately considered single-vehicle and mutiple-vehicle crashes he found that the single-vehicle accident rate decreased with annual traffic while the multiple-vehicle rate increased.

Ceder and Livneh (8) used both time-sequence analysis and cross-sectional analysis to study single- and multivehicle accidents for a series of eight roadway segments over an 8-year period. Ceder (9) expanded on this work by considering accident rates in conjunction with free-flow and congested-flow conditions. He found that the total accident rate versus hourly flow curve followed a U-shaped configuration for the free flow, which is the result of a convex downward and a convex upward curve for single- and multivehicle accidents, respectively. For congested flow, the accident rate for multivehicle accidents increases rapidly with hourly traffic flow.

Transportation Center and Department of Civil Engineering, Northwestern University, Evanston, Ill. 60201.

A TIME-SPACE PERSPECTIVE ON ACCIDENT OCCURRENCE

These studies can be considered as representing different areas in a space-time plane (see Figure 1). Smeed's research represents a very large area because he used national statistics on an annual basis (7). The use of horizontal lines indicates that the analysis was cross-sectional, comparing the accident experiences of different countries (a spatial analysis). Gwynn considered only one route but conducted his comparisons on an hourly basis in the time domain, so there are vertical lines within the domain defining his study (5). These areas are not drawn to scale but are used to illustrate how these two studies would be represented in the space-time plane.

Different types of accident studies result in different shapes in the space-time plane. Each of these shapes can be linked to particular probability distributions that describe the probability of accident occurrence. For example, with a long time period and large study section it may be reasonable to approximate the occurrence of accidents by a normal distribution. If the time and space domains are large, there is a very small likelihood of zero accidents in a time interval. The normal distribution will then have a large mean and a comparatively small variance that will make zero or negative values unlikely. This appears to be a reasonable distribution to use in this context.

The spatial and temporal aggregation required by these large time-space areas makes it difficult if not impossible to isolate the influence of driver and environmental characteristics. The normal distribution that is assumed in linear regression and analysis of variance (ANOVA) may be appropriate for

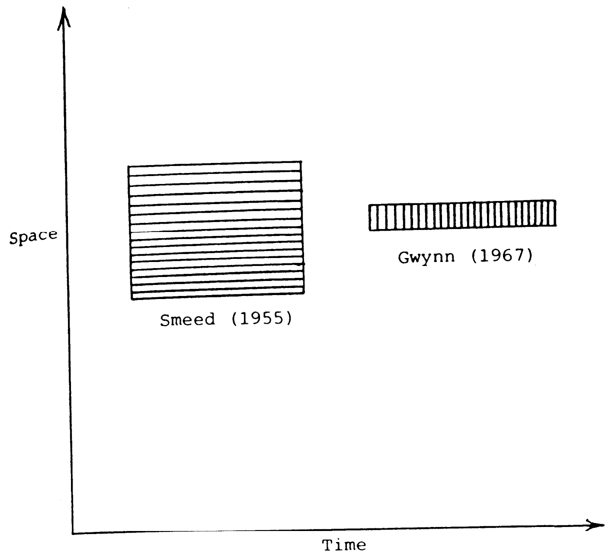

FIGURE 1 Space-time plane for highway accident study.

some specific comparisons (e.g., accident frequency on particular roadway segment) but controls for driver and vehicle factors must be carefully developed particularly for cross-sectional studies. Although vehicle and driver characteristics may be known for accidents, these attributes may be very difficult to obtain for nonaccidents.

At the other extreme of the time scale are analyses that seek to explore accident occurrence as a series of "success" or "failure" trials in which success is the safe completion of a trip and failure is an accident. Over a period of time accidents can be considered as governed by this Bernoulli process (10). At the level of individual trip success and failure trials, much detail can be retained about the driver and the vehicle. Although roadway and environment data are available at the accident scene, it is more difficult to obtain these data for the other portion of the trip. Data collection would have to be carefully managed as very large sample sizes may result.

One extension of this representation as a Bernoulli process is the use of survival theory to predict the probability of having an accident at a particular time, given that the driver has survived until that time (11). Survival theory also provides for inclusion of successful trips that are completed without an accident as exposure data. Although more research needs to be conducted to improve variable specifications, this model holds great promise in allowing improved representation of driver, vehicle, roadway, and environment in a consistent framework. A major potential disadvantage of the Bernoulli approach is that data collection and assembly may be expensive, particularly for exposure data.

Intermediate between these two representations is the Poisson distribution, which allows for discrete outcomes that are strictly nonnegative. If the time and space domains are carefully defined, and the assumption about independence of events is satisfied, the Poisson distribution may be reasonable.

OBJECTIVES

The discussion of time-space domains for accident analysis introduces the concept that different study designs can lead to different underlying probability distributions describing accident occurrence. The objective of this paper is to further explore this issue. If the true causes (and potential countermeasures) of accidents are to be identified, then statistical procedures must be used that accurately describe accident occurrence. In the remainder of the paper a model of accident occurrence that offers important advantages over conventional linear regression methods is developed and tested.

In the next section of the paper particular properties of accident occurrence that can complicate analysis are discussed. Poisson regression is suggested as a means to overcome many of these complications. The analytical structure for Poisson regression is described in a later section. Results and discussion of a model application are presented followed by conclusions.

DISCRETENESS, NONNEGATIVITY, AND HOMOSCEDASTICITY

Regression analysis has been widely applied in studies that seek to relate accident occurrence to traffic volume (8-9). Three particular properties of accident occurrence argue for care in the application of linear regression to road safety studies. These properties are the existence of relationships between the mean and variance of accident frequency, nonnegativity of the dependent variable (either accident frequency or accident rate), and occurrence of nonnormal error term distributions. Each of these issues is discussed in detail in the paragraphs that follow.

It is common to think of accident occurrence as a process that follows a Poisson or possibly a Bernoulli process (10). Both of these processes imply that the variance of accident frequency is functionally related to the mean (e.g., in Poisson processes the variance is equal to the mean). If an attempt is made to regress accident frequency by vehicle miles of travel for an accident process that is actually Poisson, the results obtained might be similar to Figure 2. Because more accidents are generally likely to occur at higher traffic volumes (due to increased conflicts), a linear positive relationship would be expected to fit through such data. It can readily be seen from Figure 2, however, that as VMT increases so does the variance of accident frequency (the dependent variable). This condition clearly violates the homoscedasticity assumption (error term has equal variance for the entire range of the predictor variables) of linear regression (12).

Violation of the assumption of equal variance of the error terms will not affect the estimated parameters; it does affect the confidence intervals of the estimators, invalidating any hypothesis tests concerning the significance of the parameters. If the objective of a study is to determine the influence that particular predictor variables have on accident occurrence, the failure to properly test for parameter significance is a serious flaw.

The use of accident rates (accidents and quantity of traffic) in the regression analyses may appear to overcome the problems with functionally related means and variances. Figure 3 shows a comparable regression line for accident rate regressed against VMT. Despite the transformation to a continuous dependent variable, one can still sketch in contours of equal accident frequency per unit time (whatever the time dimension of exposure). This estimation still results in a violation of the homoscedasticity assumption of linear regression.

Assume that accident frequency for a study section is governed by a Poisson process (mean = variance = λ) and that the frequency will increase with in-

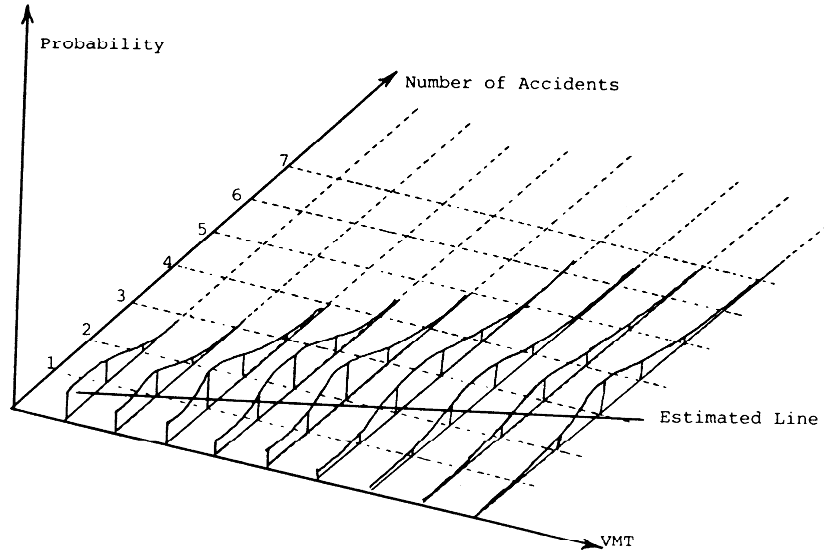

FIGURE 2 Regressing accident frequency and VMT.

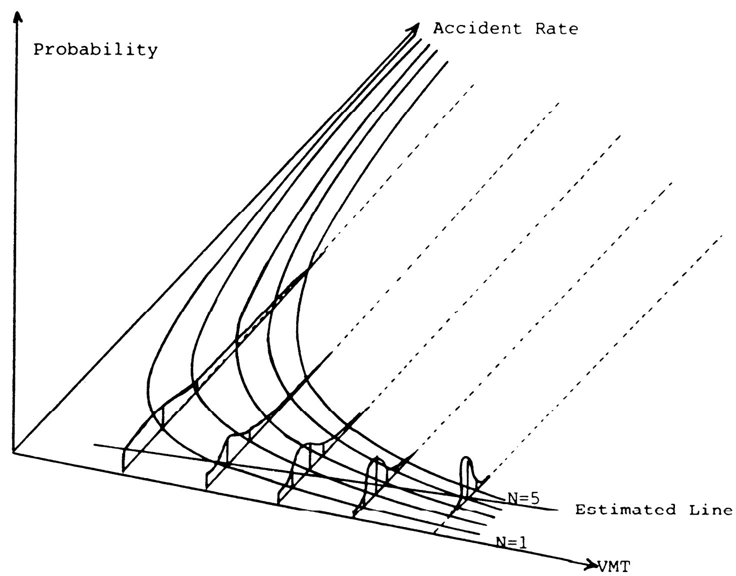

FIGURE 3 Regressing accident rate and VMT.

creasing VMT. The regression relationship represented in Figure 2 can be written as

$$y_t = ax_t + u_t \tag{1}$$

where

y_t = accident frequency in period of time, t;
x_t = VMT during time t;
u_t = error term, assumed to be distributed normal (o, σ_u^2); and
a = regression parameter.

If accident occurrence is a Poisson process, then the variance of u_t will be related to the VMT through the Poisson parameter λ, which will increase with increasing VMT. Therefore

$$\mathrm{Var}(u_t) = x_t\sigma_u^2 \tag{2}$$

This problem can be corrected by using a variance stabilizing transformation (13); specifically, dividing Equation 1 by $(x_t)^{1/2}$. This yields

$$y_t/(x_t)^{1/2} = a(x_t)^{1/2} + u_t^* \tag{3}$$

$$u_t^* = u_t/(x_t)^{1/2} \tag{4}$$

then

$$\begin{aligned}\mathrm{Var}(u_t^*) &= E[(u_t^*)^2] = E(u_t^2/x_t) = (1/x_t)E(u_t^2)\\ &= (1/x_t)(x_t)\sigma_u^2 = \sigma_u^2\end{aligned} \tag{5}$$

Therefore error terms of equal variance will result for the case when Equation 1 is divided by $(x_t)^{1/2}$. If y_t/x_t (accident rate) is regressed with x_t, the result will be narrower variance at higher levels of VMT. This is shown in Figure 3 by a smaller variance as VMT increases.

These variance-stabilizing transformations are useful, but when the response variable has been reexpressed, the predicted values are in the transformed scale. It is often necessary to convert the predicted values back to the original units. Unfortunately, applying the inverse transformation directly to the predicted values gives an estimate of the median of the distribution of the response instead of the mean (14).

The second problem, nonnegativity of accident occurrence, also imposes restrictions on the application of linear regression. The restriction is apparent in Figures 2 and 3 for both discrete and continuous dependent variables. If either regression is conducted for a set of data with high accident frequency or accident rate, respectively, then the prediction of "negative" accidents is much less likely. The requirement for either high frequency or high rate carries obvious implications for study design (particularly in the time-space plane of Figure 1). Restricting cases to those with high accident frequency may also increase study costs by requiring a larger sample of data.

A number of analytic methods are available to deal with these estimation problems. The method of least squares subject to a priori constraints can overcome the problem of negative-value prediction, but it will lead to biased estimates of model coefficients (15). Nonlinear models are also used to avoid the negative-value prediction problem, and a least squares estimation procedure based on the linearization of the nonlinear form (such as logarithm) can be applied to estimate the parameters. However, the logarithm of zero is not defined, and a zero accident observation therefore cannot be included in the investigation. One alternative for dealing with this problem is to omit the zero observations, but this is undesirable because the traffic situations in which no accidents occur are obviously important. The other alternative is to add a small number (e.g., 0.04) to all observations of the dependent variable (16). Such pretreatment of observations can greatly affect the estimation and is therefore undesirable.

A third problem occurs when the error terms are not normally distributed due to the characteristics of nonnegativity and small value of discrete dependent variable (see Figure 2). Under these conditions the correct confidence intervals will not be obtained for estimated parameters, and tests of parameter significance are again invalid.

The Poisson regression model, which assumes that the occurrence of the dependent variables follows the Poisson distribution, can effectively overcome most of the problems caused by discrete and nonnegative values of observations in normal linear regression analysis. Poisson regression techniques were used for the analyses of accidents in The Netherlands (17). Each accident was assigned to a category defined by different ranges of independent variables. For example, Hamerslag specified 4 classes for motor vehicle volume and 4 classes for motor bicycle volume, obtaining 16 categories of accidents through the combination of those two independent variables. The expected annual accident frequency for each category was assumed to be a function of its respective classes of independent variables. The number of accidents that occurred within a given time interval for each category was assumed to be Poisson-distributed with the mean equal to the predicted accident frequency for that category.

Aggregating information in the specification of the independent variables by range hinders the ability to explore the risk factors for traffic operation. Furthermore, if the number of independent variables is large, a huge sample size of data is required in order to obtain statistical power. The authors' research model is an attempt to apply the Poisson model to a more disaggregate analysis of highway accidents, to better identify some of the factors contributing to highway operating risk.

POISSON REGRESSION MODEL

The Poisson distribution was first considered in the context of regression analysis about two decades ago (18). It assumes that the dependent variables in a regression analysis are counts that follow the Poisson distribution, and that the observations are independent with the expectation as defined by the following equation:

$$E(y_{ij}) = f(\underline{x}_i, \underline{\beta}) \qquad i = 1,2 \ldots\ldots\ldots, n. \qquad j = 1,2 \ldots\ldots\ldots, m_i. \tag{6}$$

where $\underline{x}_i = (x_{i1}, x_{i2}, \ldots\ldots, x_{ik})$ is the ith set of values of the k independent variables, m_i is the number of replications of the ith experimental condition, $\beta = (\underline{\beta}_1, \beta_2, \ldots\ldots, \beta_p)'$ is a p-dimensional vector of unknown parameters, and (y_{ij}) is a particular realization of the experiment (15). It is further assumed that some general form of the model is known and $f(\underline{x}, \underline{\beta})$ is a differentiable function of $\underline{\beta}$. Then n values of the independent variables are selected by the experimenters or specified by the situation. The number of n is supposed to be sufficiently greater than p to ensure estimability of the parameters. Three different methods are available to estimate the parameters of the Poisson regression model. They are maximum likelihood, weighted least squares, and minimum chi-square estimation. Maximum likelihood estimation has been widely accepted in past applications because of its convenience.

The occurrence of highway accidents can be reasonably described by the nonstationary Poisson process if the study system (or area) is adequately selected. According to the basic assumptions of the Poisson process, it is assumed that the number of accidents occurring within each observed time interval is independent, with the expectation defined as in Equation 6. This expectation of the number of accidents in each time interval is a function of traffic volume, road and weather conditions, and so forth. Hence the expected values of accidents are different from time interval to time interval, and this is the so-called nonstationary Poisson process. The model is set as

$$\lambda_i = f(\underline{\beta}, x_i) \tag{7}$$

where

- λ_i = expected value of accident frequency for ith time interval,
- $\underline{\beta}$ = the vector of parameters to be estimated, and
- $\underline{X}_i$ = the vector of independent variables for ith time interval.

The probability of k accidents occurring in t intervals is represented as

$$P_i(k) = (e^{-\lambda_i t})(\lambda_i t)^k / k! \tag{8}$$

However, because only the accidents occurring in each time interval (t=1) are considered, Equation 8 becomes

$$P_i(k) = e^{-\lambda_i} (\lambda_i)^k / k! \tag{9}$$

Then a set value of $\hat{\underline{\beta}}$ that maximizes the following likelihood value (L) is sought:

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=0}^{\infty} P_i(k)^{D_{ik}} \qquad (10)$$

where D_{ik} is the dummy variable for the number of accidents that occur in the ith time interval:

D_{ik} = 1, if k accidents occurred in ith time interval,

D_{ik} = 0, otherwise.

For convenience, a logarithm transformation of Equation 10 is taken and called the log-likelihood value $[LL(\hat{\underline{\beta}})]$:

$$LL(\hat{\beta}) = \sum_{i=1}^{n} \sum_{k=0}^{\infty} D_{ik} \log [P_i(k)] \qquad (11)$$

LL(c) is also defined as the log-likelihood value of the model in which only the constant term is used. The value of $2[LL(\hat{\underline{\beta}})-LL(c)]$ is distributed as χ^2-distribution with p-1 degrees of freedom. It is a statistic for testing the significance of all explanatory variables included in the model. ρ^2, defined as $1-[LL(\hat{\underline{\beta}})/LL(c)]$, is an informal goodness-of-fit measure and is analogous to R^2 used in regression.

DATA COLLECTION AND MODEL FORMULATION

Data Collection

The data used in this study were collected on the Indiana Toll Road in 1978. The Indiana Toll Road is an east-west road 157 mi long. It transverses mostly open flat country. There are no steep grades or sharp horizontal curves. Daily VMT data were derived from the toll collection system. All drivers entering the Indiana Toll Road receive a ticket that records the designation of the particular interchanges of entry and the date on which the vehicle entered the toll road. The toll is collected when the vehicle leaves the tollway, and the vehicles are then classified according to the toll schedules' list of vehicle classes. Because a vehicle can only enter and leave the tollway at a limited number of interchanges, these facilities are closed systems where VMT can be easily calculated and recorded by mechanized card-reading procedures. Thus, the precise daily automobile and truck VMTs are available.

Data describing all toll road main line accidents were obtained from records of the Indiana State Police Toll Road Headquarters. Accidents occurring at toll booths, access roads, service areas, and ramps were excluded because they are likely to be influenced by geometric design and other operational characteristics that are site-specific. By studying only main line accidents, the authors hoped to obtain a clearer relationship between accidents and exposure for a well-designed, four-lane freeway facility. After screening non-main line accidents, the data set included more than 700 accidents and 1,023 vehicle involvements (19).

The weather data were obtained from the National Oceanic and Atmospheric Administration's Environmental Data and Information Service at the National Climatic Center in Asheville, North Carolina. There are six stations recording hourly precipitation and amount of daily snowfall along the Indiana Toll Road. The toll road is within 5 mi of all the stations, which are roughly evenly spaced. The hours of snow and hours of rain were derived by Delleur (19); an average of the weather conditions of those six survey stations is used as the hours of snow and hours of rain for the toll road overall.

Independent Variables

A regression model does not imply a cause-and-effect relationship between the variables. To establish causality, the relationship between the regressors and the response must have a basis outside the sample data; for example, the relationship may be suggested by theoretical considerations. Regression analysis can aid in confirming a cause-effect relationship, but it cannot be the sole basis.

Traffic volume and traffic composition affect traffic speed, variation of vehicle traveling speeds, and drivers' psychological condition. For example, automobile drivers may feel uncomfortable when they join a traffic stream that has a high truck volume. Hence, the increase of traffic volume will not only increase the number of accidents because of more exposure, but it will also increase traffic conflict and friction. VMT is used to represent the traffic volume in the study system. Daily traveling miles of automobiles, small trucks, and large trucks are separated not only for exposure considerations, but also for distinguishing their effects on different accident patterns. Small trucks include six-tire vehicles with two axles, commercial vehicles with three axles, and two-axle tractors with one-axle trailer.

The second factor tested in this model is the weather condition, which influences the friction of roadway pavement and driver's sight distance. These effects will affect the safety of high-speed operation on the highway. The hours of snowing and raining in the study system are considered to reflect the effect of daily weather condition on daily accident occurrence. An average of the data collected from the survey stations along the toll road is used to represent the daily weather condition.

The last regressor considered in this study is the effect of different driving populations on accident occurrence. Travel is derived by people's activities, which are generally controlled during weekdays by work trips and related travel. These drivers are likely to be frequent travelers of the toll road, familiar with its relatively high mix of automobiles and trucks. Weekend travelers may be less frequent users who are less able to cope with traffic conditions on the road. In order to help capture the influence, if any, of different driving populations, a dummy variable--labeled WEND--is included in the model.

Functional Form of the Model

There is no reason to prefer one functional form over any other for Equation 7. A linear additive form was initially tested but failed to result in valid model estimates. It appeared that some sets of possible parameter estimates caused $\hat{\lambda}$ to be negative, violating the assumed conditions of the Poisson distribution. A multiplicative specification was also tested and yielded valid parameter estimates. The multiplicative form of Equation 7 is as follows:

$$\lambda = \beta_0 (\mathrm{VMTa})^{\beta_1} (\mathrm{VMTlt})^{\beta_2} (\mathrm{VMTst})^{\beta_3} (1+\mathrm{HSNOW})^{\beta_4} (1+\mathrm{HRAIN})^{\beta_5} (1+\mathrm{WEND})^{\beta_6} \qquad (12)$$

where

λ = expected accident frequency per day;
VMTa = daily VMT of automobiles (10^6 vehicle miles);
VMTlt = daily VMT of large trucks (10^6 vehicle miles);
VMTst = daily VMT of small trucks (10^6 vehicle miles);
HSNOW = hours of snow in the study system;
HRAIN = hours of rain in the study system; and
WEND = dummy variable for weekend; WEND=1 for weekend, and WEND=0 otherwise.

One has been added to each of the last three predictor variables to prevent zero values for estimated λ's, which would result in the logarithm of zero (which is undefined) occurring in Equation 6. Notice that this formulation does allow for zero values of the dependant variable so that it is not subject to the criticisms described in the section on Discreteness, Nonnegativity, and Homoscedasticity.

MODEL RESULTS AND INTERPRETATION

Preliminary empirical results indicated that the parameters of the weekend and small truck VMT were not significant. The dummy variable, WEND, was removed from the original model in order to improve the model structure. A χ^2-test, revealed that deleting weekend caused no significant difference in the explanatory ability of the restricted model compared to the original model. The small truck VMT also had no significant effect on accident occurrence. The small truck VMT was small relative to automobile VMT and large truck VMT, and it also was positively correlated with the large truck VMT. In order to deal with the issue of collinearity between independent variables, the small truck and large truck VMTs were combined into one variable, truck VMT.

Therefore, only four independent variables were included in the model: (a) automobile VMT, (b) truck VMT, (c) hours of snow, and (d) hours of rain. The empirical results of the final model are given in Table 1, along with the log likelihood and goodness-of-fit measures discussed in the section titled Poisson Regression Model.

In addition to the pooled model, which includes all accidents, several additional models are estimated for different types of collisions. Separate models are estimated for single-vehicle crashes (both automobile and truck) and for two-vehicle crashes differentiated as automobile-automobile, truck-truck, and truck-automobile. In addition, separate models were estimated for single vehicle and multiple vehicle crashes. Crashes with three or more vehicles were extremely rare, so more than 95 percent of the accidents were included in these categories. The different models were separately estimated to determine if VMT and weather conditions may have had different effects on different types of vehicle crashes.

Pooled Model

The estimates for the pooled model indicate that all parameters are significant except for hours of rainfall. The χ^2-test for the entire model strongly rejects the null hypothesis that the full model has explanatory power equal to that of the model with the constant term only. Consistent with previous results (2), hours of snowfall is strongly positively associated with accident occurrence as are automobile and truck VMT.

Although the pseudo goodness-of-fit measure is small (ρ^2 = .06), this is an indication of the additional variation in accident frequency explained by the four predictors compared to the constant term alone. The magnitude of the additional variation explained by the predictor variables is not incon-

TABLE 1 Empirical Results for Poisson Regression Models

Parameter (Accident Type)		Pooled Model	Single Auto	Single Truck	Auto-Auto	Truck-Truck	Auto-Truck	Auto Involved	Truck Involved	Single Vehicle Accident	Multiple Vehicle Accident
Constant	β_0	1.775	0.572	0.481	0.128	0.293	0.359	1.038	1.117	1.047	0.735
	t	(17.20)	(10.33)	(8.48)	(5.26)	(5.80)	(7.25)	(13.74)	(12.54)	(13.43)	(10.75)
VMTauto (10^6)	β_1	0.255	0.513	-0.066	0.984	-0.929	0.146	0.502	-0.135	0.300	0.183
	t	(3.49)	(-4.99)	(-0.44)	(4.25)	(-3.84)	(0.79)	(5.40)	(-1.32)	(3.19)	(1.57)
VMTtruck (10^6)	β_2	0.229	-0.059	0.674	-0.412	1.769	0.556	0.005	0.751	0.184	0.311
	t	(2.48)	(-0.40)	(3.12)	(-1.68)	(4.41)	(2.28)	(0.05)	(4.94)	(1.56)	(2.07)
Snow (hrs)	β_3	0.626	0.569	0.683	0.741	0.377	0.673	0.634	0.624	0.617	0.636
	t	(9.90)	(5.32)	(5.41)	(4.22)	(1.85)	(4.47)	(8.12)	(7.18)	(7.55)	(6.39)
Rain (hrs)	β_4	0.025	0.180	0.097	-0.186	-0.277	-0.252	0.037	-0.076	0.150	-0.244
	t	(0.39)	(1.93)	(0.74)	(-0.87)	(-1.06)	(-1.35)	(0.48)	(-0.77)	(1.98)	(-1.97)
Average Daily Accidents		1.995	0.789	0.466	0.230	0.178	0.332	1.351	0.975	1.255	0.740
Number of Accidents in 1982		728	288	170	84	65	121	493	365	458	270
LL(c)		-714.29	-455.46	-339.00	-222.19	-184.21	-278.15	-605.67	-516.44	-566.55	-451.57
LL($\hat{\beta}$)		-668.77	-434.41	-321.82	-200.16	-164.81	-266.01	-565.37	-478.63	-539.42	-428.06
$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(c)}$		0.064	0.046	0.051	0.099	0.105	0.044	0.067	0.073	0.048	0.052
$2(LL(\hat{\beta}) - LL(c))$		91.0	42.1	34.4	44.1	38.8	24.3	80.6	75.6	54.3	47.0

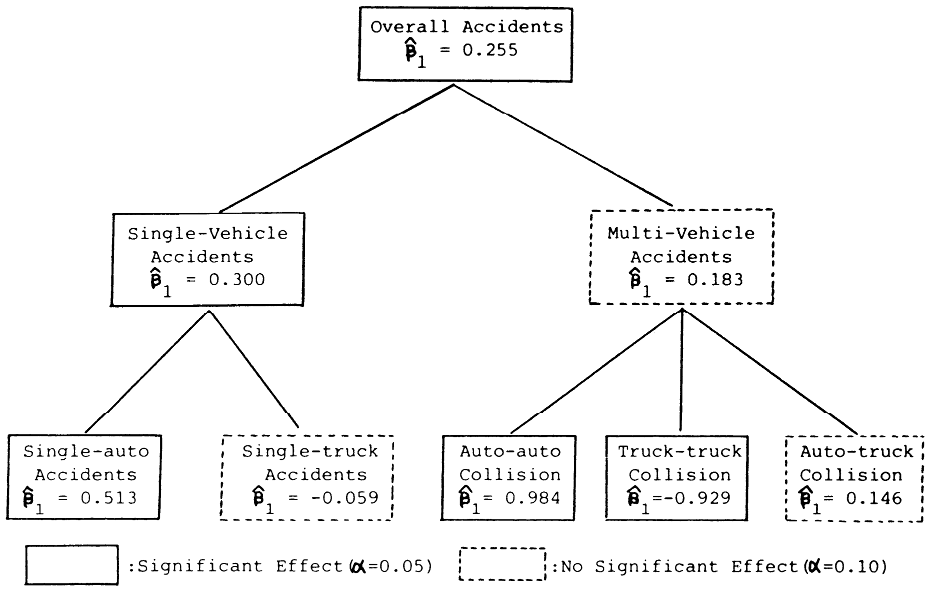

FIGURE 4 Effects of automobile VMT on the accident occurrence for different accident patterns.

sistent with results at disaggregate models in the travel demand literature. It must be remembered that the dependent variable is daily accident frequency. A significant amount of random variation might be expected with such a variable.

The summaries in Table 1 include values for LL(c), LL($\hat{\beta}$), ρ^2 and 2[LL($\hat{\beta}$) - LL(c)] for each of the models of the individual accident types. These statistics can be used to test the improvement in model fit that is obtained when a more detailed analysis is conducted (i.e., when one moves from a pooled model to single-vehicle crashes to separate single automobile and single truck collisions as is shown in Figure 4). The findings were consistent: the detailed models always achieved a statistically significant improvement in goodness of fit compared with the less-detailed models.

Effect of Automobile VMT

Figure 4 is an overview of the parameters for automobile VMT for each of eight separate but related models. The tree structure sequentially separates accidents into the more detailed categories. By estimating this sequence of models and comparing the significance and magnitude of a parameter, the influence of automobile VMT on different types of accidents can be determined. In addition to the parameter value, the figure also indicates statistical significance. As would be expected, automobile VMT is significant for the pooled model (overall accidents) as well as single-automobile and automobile-automobile collisions.

In order to better understand the interaction of the types of collisions, Figure 5 was constructed.

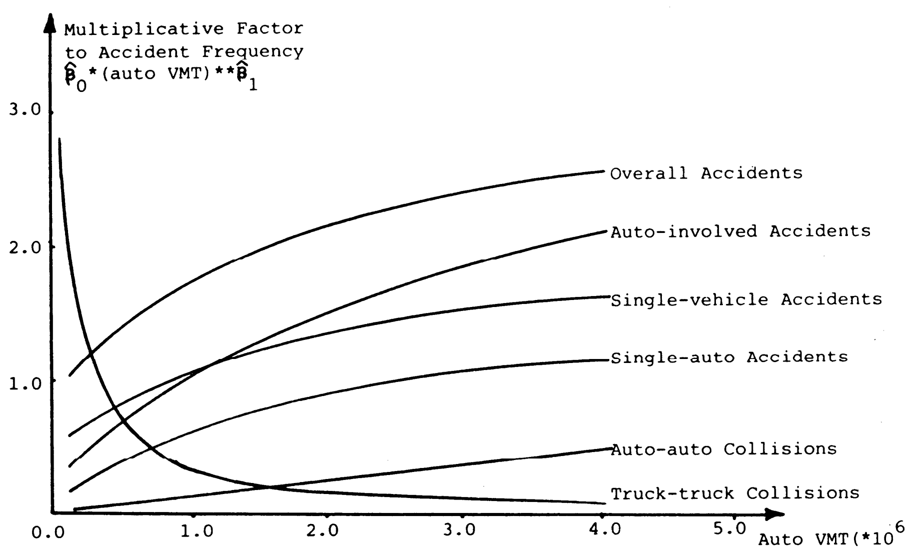

FIGURE 5 Effects of automobile VMT on accident occurrence within the range of available data.

For comparison purposes, the constant term, $\hat{\beta}_0$, is combined with the effect of automobile VMT and called multiplicative factor. The number of overall accidents increases at a decreasing rate ($0 < \hat{\beta}_1 < 1.0$) as the automobile VMT increases. This means that increases in automobile traffic will increase the number of overall accidents when the other factors are fixed, but the accident rate, which can be determined by (VMTauto)**($\hat{\beta}_1$-1), will decrease.

This increase of overall accidents is mainly attributed to the increase of automobile-involved accidents, especially the single-automobile accidents.

From Figure 5, it can be seen that the curves for single-vehicle accidents and single-automobile accidents are parallel over the range of available automobile VMT data. This implies that the automobile VMT has no effect on the single-truck accidents. The number of single-automobile accidents increases at a decreasing rate, whereas the number of automobile-automobile collisions increases approximately linearly ($\hat{\beta}_1 \doteq 1.0$). Hence, the proportion of automobile-automobile collisions to automobile-involved accidents will increase as the automobile VMT increases.

The number of truck-truck collisions sharply decreases when the automobile VMT increases. This decrease might be expected to be compensated by an increase in truck-automobile collisions. However, neither a significant decrease of single-truck accidents nor a significant increase of truck-automobile collisions can be found. The effect on total multiple vehicle accidents is not significant because of the compensating effects on truck-truck and automobile-automobile crashes (this is also shown in Figure 4). This is a clear example of how this market segmentation approach can be used to gain information about the influence of predictor variables on different types of accidents.

Effect of Truck VMT

Truck VMT also has a significant effect on overall accidents (see Figures 6 and 7). The number of overall accidents increases at a decreasing rate ($0 < \hat{\beta}_2 < 1.0$) as the truck VMT increases (Figure 6). This increase of overall accidents is mainly attributed to the significant increase of truck-involved accidents, which include the single-truck, truck-truck, and truck-automobile accidents. The single-truck accidents and truck-automobile collisions increase at a decreasing rate, whereas the truck-truck collisions increase at an increasing rate. From Figure 7, it can be seen that the number of truck-truck collisions will occupy a significant proportion of truck-involved accidents when truck traffic is high. The number of automobile-automobile collisions marginally decreases as the truck VMT increases (Figure 6). The increase of truck-automobile collisions and decrease of automobile-automobile collisions when the truck VMT increases verify the hypothesis that automobile-automobile collisions shift to truck-automobile collisions as truck VMT increases.

Effects of Environmental Variables

The hours of snow have a significant effect on accident occurrence for all the accident patterns. The values of the parameter of snow hours are similar for all the accident patterns except the truck-truck collisions. The parameter indicates that more snow hours will increase accident occurrence at a decreasing rate ($< \hat{\beta}_3 < 1.0$), and the magnitude of this effect is quite similar for all the accident patterns except truck-truck collisions. The hours of snow have a lesser effect on truck-truck collisions than on the other accident patterns.

The hours of rain have a significant effect on single-automobile accidents, single-vehicle accidents, and multivehicle accidents. As the hours of rain increase, single-vehicle accidents increase at a decreasing rate, whereas multivehicle accidents decrease at a decreasing rate. However, overall accidents do not significantly increase. It appears that increases in rain hours tend to shift multivehicle accidents to single-vehicle accidents. The increase of single-vehicle accidents results from an increase in single-automobile accidents. In general, rainfall has much less of an effect on accident occurrence than snowfall.

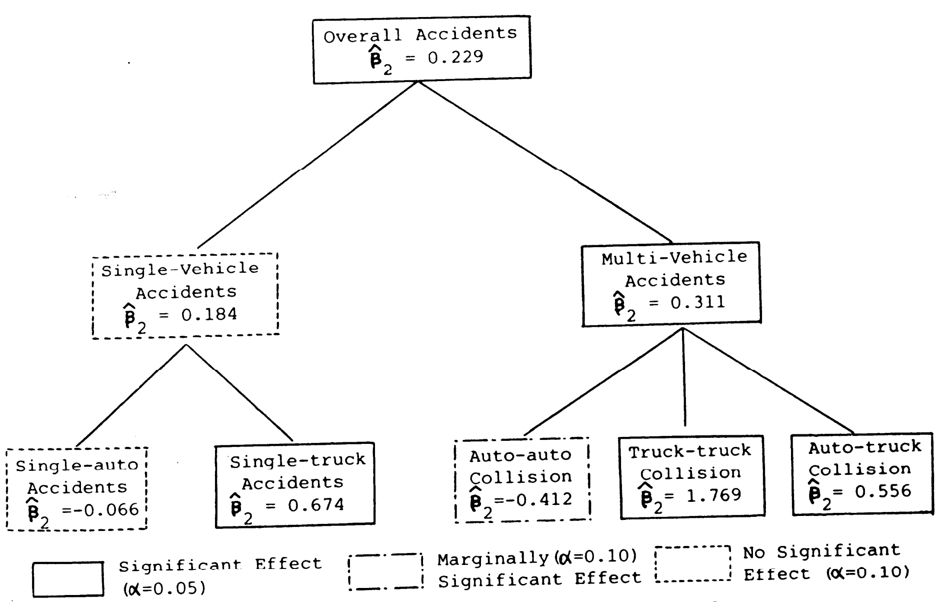

FIGURE 6 Effects of truck VMT on the accident occurrence for different accident patterns.

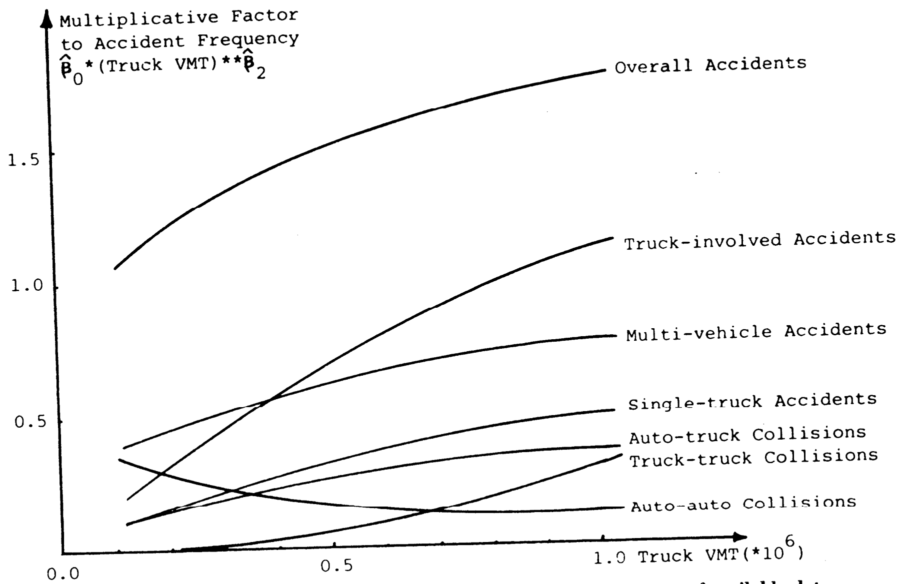

FIGURE 7 Effects of truck VMT on accident occurrence within the range of available data.

SUMMARY AND CONCLUSIONS

The analysis of highway accidents and identification of factors contributing to their occurrence is a complex process. A time-space framework is presented to facilitate a review of the literature and introduce the use of various probability distributions to model accident occurrence.

The normal distribution, which underlies traditional linear regression and hypothesis testing methods, should be used with caution because of problems associated with nonnegativity and error terms with unequal variance. If the underlying accident process is one in which the mean accident frequency is functionally related to the variance (e.g., Poisson distribution), parameters in a linear regression model will be unbiased but will have incorrect confidence limits. If the objective of the regression is to identify factors that significantly affect accident occurrence, incorrect confidence limits invalidate hypothesis tests of parameter significance--a serious shortcoming. Regressing accident rates rather than accident frequency may still result in unequal error variances, particularly when the underlying process is Poisson.

Poisson regression applied directly to accident data is proposed as a method to overcome many of these shortcomings. A Poisson regression model is applied to daily accident, travel mileage, and environmental data from the Indiana Toll road. Market segmentation is used to study whether VMT and weather conditions have different effects on different types of vehicle crashes. The models reveal that automobile and truck accidents are directly related to automobile and truck travel (as expected). As truck VMT increases, there is also a marginal reduction in automobile-automobile collisions and an increase in automobile-truck collisions. Snow strongly affects all accident types, whereas rainfall primarily increases the mean automobile accident frequency and has no effect on trucks.

Poisson regression has superior statistical properties for many potential applications to highway safety. In addition, it can be used with generally smaller sample sizes than linear regression. In conjunction with the use of segmentation, it can yield important insights about the significance of factors in accident occurrence.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by Yellow Freight Systems, Inc., United Parcel Service Foundation, and the Gar Foundation of Roadway Express, Inc.

REFERENCES

1. R. Chapman. The Concept of Exposure. Accident Analysis and Prevention, Vol. 5, No. 2, 1973, pp. 95-110.
2. P.P. Jovanis and J.B. Delleur. Exposure-Based Analysis of Motor Vehicle Accidents. In Transportation Research Record 910, TRB, National Research Council, Washington, D.C., 1984, pp. 1-7.
3. D.M. Belmont. Effect of Average Speed and Volume on Motor Vehicle Accidents on Two-Lane Tangents. HRB Proc., Vol. 32, 1953, pp. 383-395.
4. W. Leutzbach. Traffic Accidents and Traffic Flow on the Frankfurt-Mannheim Motorway--Investigation of the Relationship between Traffic Accidents and Traffic Flow (in German). Hessische Ministerium fur Wirtschaft und Verkehr, March 1966.
5. D.W. Gwynn. Relationship of Accident Rates and Accident Involvements with Hourly Volume. Traffic Quarterly, Vol. 21, No. 3, July 1967, pp. 407-418.
6. W.T. Baker and D.W. Gwynn. Relationships of Accident Rates with Hourly Traffic Volumes. Division of Research Evaluation, New Jersey Department of Transportation, Newark, 1968.
7. R.J. Smeed. Accident Rates. International Road Safety and Traffic Review, Vol. 3, No. 2, 1955, pp. 30-40.
8. A. Ceder and M. Livneh. Relationships Between Road Accidents and Hourly Traffic Flow--I, Analysis and Interpretation. Accident Analysis and Prevention, Vol. 14, No. 1, 1982, pp. 19-34.

9. A. Ceder. Relationships Between Road Accidents and Hourly Traffic Flow--II, Probabilistic Approach. Accident Analysis and Prevention, Vol. 14, No. 1, 1982, pp. 35-44.
10. W.L. Cresswell and P. Froggatt. The Causation of Bus Driver Accidents: An Epidemiological Study. Oxford University Press, London, England, 1963.
11. P.P. Jovanis and H. Chang. A Disaggregate Model of Highway Accident Occurrence. Forthcoming, Accident Analysis and Prevention, 1986.
12. C.K. Ruijgrok and P.G. Van Essen. The Development and Application of Disaggregate Poisson Model for Trip Generation. Presented at 59th Annual Meeting of the Transportation Research Board, Washington, D.C., 1980.
13. H.H. Kelejian and W.E. Oates. Introduction of Econometrics. 2nd ed., Harper and Row, New York, 1981.
14. D.C. Montgomery and E.A. Peck. Introduction to Linear Regression Analysis. John Wiley and Sons, Inc., New York, 1982.
15. E.L. Frome et al. Regression Analysis of Poisson-Distributed Data. Journal of American Statistical Association, Vol. 68, No. 344, Dec. 1973, pp. 935-940.
16. R.A. Lavette. Development and Application of a Railroad-Highway Accident--Prediction Equation. In Transportation Research Record 628, TRB, National Research Council, Washington, D.C., 1977, pp. 12-19.
17. R. Hamerslag, J.P. Roos, and M. Kwakernaak. Analysis of Accidents in Traffic Situations By Means of Multiproportional Weighted Poisson Model. In Transportation Research Record 847, TRB, National Research Council, Washington, D.C., 1982, pp. 29-36.
18. F.A. Haight. Handbook of the Poisson Distribution. John Wiley and Sons, Inc., New York, 1967.
19. J.B. Delleur. Analysis of Truck and Auto Accident Experience Based on Environmental Driving Condition and Travel Exposure. Masters thesis, Department of Civil Engineering, Northwestern University, Evanston Ill., 1983.

The findings and conclusions as well as any errors and omissions are solely the responsibility of the authors, not the project sponsors.

Publication of this paper sponsored by Committee on Traffic Records and Accident Analysis.