Accid. Anal. & Prev. Vol. 21, No. 5, pp. 445–458, 1989
Printed in Great Britain.

0001-4575/89 $3.00 + .00

DISAGGREGATE MODEL OF HIGHWAY ACCIDENT OCCURRENCE USING SURVIVAL THEORY

PAUL P. JOVANIS
Civil Engineering Department, University of California–Davis, Davis, CA 95616, U.S.A.

and

HSIN-LI CHANG
Department of Transportation Engineering and Management,
National Chiao Tung University, Hsinchu, Taiwan, Republic of China

(*Received* 22 *July* 1985; *in revised form* 1 *March* 1989)

Abstract—The analysis of discrete accident data and aggregate exposure data frequently necessitates compromises that can obscure the relationship between accident occurrence and potential causal risk components. One way to overcome these difficulties is to develop a model of accident occurrence that includes accident and exposure data at a mathematically consistent disaggregate level. This paper describes the conceptual and mathematical development of such a model using principals of survival theory. The model predicts the probability of being involved in an accident at time t given that a vehicle has survived until that time. Several alternative functional forms are discussed including additive, proportional hazards and accelerated failure time models. Model estimation is discussed for the case in which both accident and nonaccident trips are included and for the case with only accident data. As formulated, the model has the distinct advantage of being able to consider accident and exposure data at a disaggregate level in an entirely consistent analytic framework. A conditional accident analysis is undertaken using truck accident data obtained from a major national carrier in the United States. Model results are interpretable and generally reasonable. Of particular interest is that segmenting accidents in several categories yields very different sets of significant parameters. Driver service hours seemed to most strongly effect accident risk: regularly scheduled drivers who take frequent trips are likely to have a reduced risk of an accident, particularly if they have a longer (greater than eight) number of hours off-duty just prior to a trip.

1. INTRODUCTION

Background

A study of accident occurrence alone is generally not sufficient to obtain a complete understanding of accident risk. This is because the occurrence of accidents, as reflected in accident reports, for example, must be compared to the number of opportunities available to be involved in an accident. Some representation of these opportunities is commonly referred to as exposure to accident risk (shortened generally to exposure). Exposure has been formally defined as the amount of or opportunity for accidents that the driver or traffic system experiences (Chapman 1973). If we know exposure, we can differentiate high accident occurrence due to high risk from high accident occurrence due to high exposure.

Different exposure measures have been used depending on the type of accident under investigation. Examples include vehicle miles of travel, ton-miles, passenger miles, vehicle registrations, and population. A major problem in combining accident data with exposure is that accidents are discrete events. Exposure data are generally more aggregate, typically based upon measured or estimated daily, weekly, monthly, or often yearly travel. A fundamental dilemma in studies of accident occurrence is how to combine exposure and accident data in a meaningful and consistent way so that the contribution of individual factors to accident risk can be identified.

Research objective

The analysis of discrete accident data and aggregate exposure data frequently necessitates compromises that can obscure the relationship between accident occurrence and causal risk components. One way to overcome these difficulties is to develop a model of accident occurrence that includes accident and exposure data at a mathematically consistent disaggregate level. This paper describes the conceptual and mathematical development of such a model using principals of survival theory. The model is then applied to the problem of motor carrier highway safety.

Literature review

A variety of statistical approaches have been applied to studies of highway accident occurrence. The simplest is the use of the mean and variance of the accident involvement rates, which is undertaken to test the equality of accident risk between different exposure groups (e.g. Foldvary 1979; Meyers 1981; Jovanis and Delleur 1982). Linear regression has been widely used in safety studies. The dependent variable may be the number of accidents or an accident rate (per mile or million miles). The risk components, assigned as independent variables, include travel speed (Hall and Dickinson 1974; Lavette 1977), traffic volume (Oppe 1979; Ivey et al. 1981; Ceder and Livneh 1982), and weather and vehicle type (Jovanis and Delleur 1983). The automatic interaction detection (AID) technique has been used to categorize explanatory variables for different exposure segments (Snyder 1974; Cleveland and Kitamura 1978). A number of studies have discussed the advantages of using Poisson Regression as an alternative to least squares (Ruijgrok and Van Essen 1980; Hammerslag et al. 1982; Montgomery and Peck 1982; Jovanis and Chang 1986). Additional multivariate analysis techniques have been used in studies by Koornstra (1969), Hakkinen (1979), and Chirachavala et al. (1984).

The common denominator in all of the referenced studies is the search for statistical links between accident occurrence (or the risk of having an accident) and potential causal factors (referred to here as risk components). All of the studies used accident data; those that included exposure always included it at an aggregate level.

2. A CONCEPTUAL MODEL OF ACCIDENT OCCURRENCE

A starting point in the conceptual structure of the model is to consider the probability of an accident as a driver seeks to complete a trip. A driver faces many risk components when operating a vehicle on a roadway system. The risk components vary with the characteristics of the driver, vehicle, roadway, and the driving environment (Institute of Traffic Engineers 1976). The probability of being involved in an accident varies throughout a trip depending on the levels of the risk components operating at any time during a trip. High levels of a risk component can be thought of as increasing the probability of an accident. Risk factors may be constant (e.g. type of vehicle), situational, (e.g. traffic volume, road curvature), or time-related (e.g. driver fatigue).

When an accident occurs, it may be difficult to identify the levels of the risk components that may have contributed to the crash. While this is a useful conceptualization, at the empirical level we may be able to do no more than identify variables that are associated with accident occurrence without knowing for certain whether they "caused" the crash. Risk components that contribute to accident occurrence should be more readily identified if this conceptualization of individual vehicle risk can be formulated in a consistent mathematical framework. In a manner consistent with the positive guidance literature, one can think of an accident as a "system failure." The probability of a system failure is affected by the levels of the risk components at any time during the trip. The following section discusses alternative models of system failure and describes a particular model that is adapted for highway safety analysis.

3. FORMULATION OF SYSTEM HAZARD

Background

Several approaches are available to describe the failure of one operating system facing a variety of risk components. Those approaches vary from simple to complex according to the assumptions made about the relationship between risk components and system failure. Here, we briefly review those approaches and explore whether they are realistic as applied to road accidents.

Risk components can be assumed to operate independently. The operating system fails whenever the first failure of a risk component occurs. This model is called an independent competing risk model. Since failure of one risk component is enough to cause system failure, the risk components can be thought of as "competing" to make the system fail. If one observes the system, one can readily determine which component caused the system to fail but can make no statement about how close other components are to failure. Importantly, other components are assumed to have no effect.

Independent competing risk models have been applied to study the reliability of electrical equipment in which the failure of any part will shut down the system (e.g. Barlow and Proschan 1981). It seems inappropriate to assume that a "failed" single risk component causes a traffic accident. The notion is contrary to extensive investigations of crashes (Treat et al 1977) which clearly indicate that most traffic accidents have multiple contributing causes. Further, many of the risk components that can affect the probability of an accident are difficult to define as failed (e.g. weather or roadway conditions).

A second model of system failure assumes that each individual risk component contributes some hazard to the operating system depending on its level. The hazards are accumulated and the operating system fails if the cumulative hazard reaches a threshold. This threshold hazard model is good for systems in which the levels of all risk components can be easily measured throughout the survival process and for which each individual risk component has some residual effect on system failure. The threshold hazard model has been used to study the fatigue failure of materials under repeated impacts, use, or vibrations (e.g. Lieblein and Zelen 1956).

The threshold hazard model seems appropriate to describe the cumulative nature of drivers' fatigue during continuous driving. However, some highway risk components are not cumulative but clearly situational, such as a curve with poor sight distance, or constant (e.g. cargo weight).

Last, the system hazard can be expressed by the hazard of all the risk components, but no cumulative hazard is involved. That is, the probability of failure at any time is determined by the total hazard contributed by the level of each risk component at that moment. This model is called a latent system hazard model because the cause of failure is not specifically known and each component has some latent effect on the risk system. Latent system hazard models have been widely used in biomedical studies to detect the effect of specific medical treatments on the lifetimes of observed patients (e.g. Cox 1972; Cox and Oakes 1984; Kalbfleisch and Prentice 1973; Holt 1978).

A major advantage of applying the latent system hazard model to accident occurrence is to overcome the complicated issue of defining the failure of risk components and identifying the specific cause(s) of accidents. However, we can still capture the contribution of related risk components to accident involvement by comparing accident trips and nonaccident trips. Further, the cumulative hazard of some risk components (e.g. fatigue of driver due to continuous driving) can be considered by specifying the system hazard as a function of time. From the viewpoint of accuracy and applicability, the concept of latent system hazard seems appropriate to traffic safety studies. The study objects are not individuals that receive medical treatments but motor vehicles operating on a highway system. Instead of comparing the survival times of patients who are treated and untreated, the model can estimate the probability of having an accident at any time given survival until that time.

Mathematical formulation

Let T be a nonnegative random variable representing the lifetimes of individual vehicle trips in some population. Let $f(t)$ denote the probability density function of T and let the distribution function be:

$$F(t) = \Pr(T < t) = \int_0^t f(x)\, dx. \tag{1}$$

The probability of an individual vehicle surviving till time t is given by the survival function:

$$S(t) = \Pr(T \geq t) = \int_t^\infty f(x)\,\mathrm{d}x. \tag{2}$$

Note that $S(t)$ is a monotone decreasing continuous function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. The hazard function $h(t)$ is then defined as:

$$h(t) = \lim_{\Delta t\to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{3}$$

The hazard function thus represents the probability of failure at time t given that the vehicle survives until time t. It is also useful to define the cumulative hazard as:

$$H(t) = \int_0^t h(x)\,\mathrm{d}x. \tag{4}$$

The lifetime of an individual trip is affected by risk components; thus, the hazard function is generally represented as $h(t|\mathbf{X})$, where $\mathbf{X}$ is a vector of risk components which may contribute to accident occurrence.

Several forms for the hazard function have been discussed in the biostatistics literature (Aranda-Ordaz 1983; Prentice and Kalbfleisch 1979; Lawless 1982), including additive, proportional hazards, and accelerated failure time models. The general form of the additive model is:

$$h(t|\mathbf{X}) = h_0(t) + K(\mathbf{X}), \tag{5}$$

in which $h_0(t)$ is the underlying hazard when $K(\mathbf{X}) = 0$. Whatever the functional form of $K(\mathbf{X})$, eqn (5) implies that the risk components specified in $K(\mathbf{X})$ are totally independent of the other risk components that are implicitly included in $h_0(t)$ (see Fig. 1a). This assumption seems unrealistic as applied to highway safety studies.

The multiplicative model can be represented as:

$$h(t|\mathbf{X}) = h_0(t)G(\mathbf{X}), \tag{6}$$

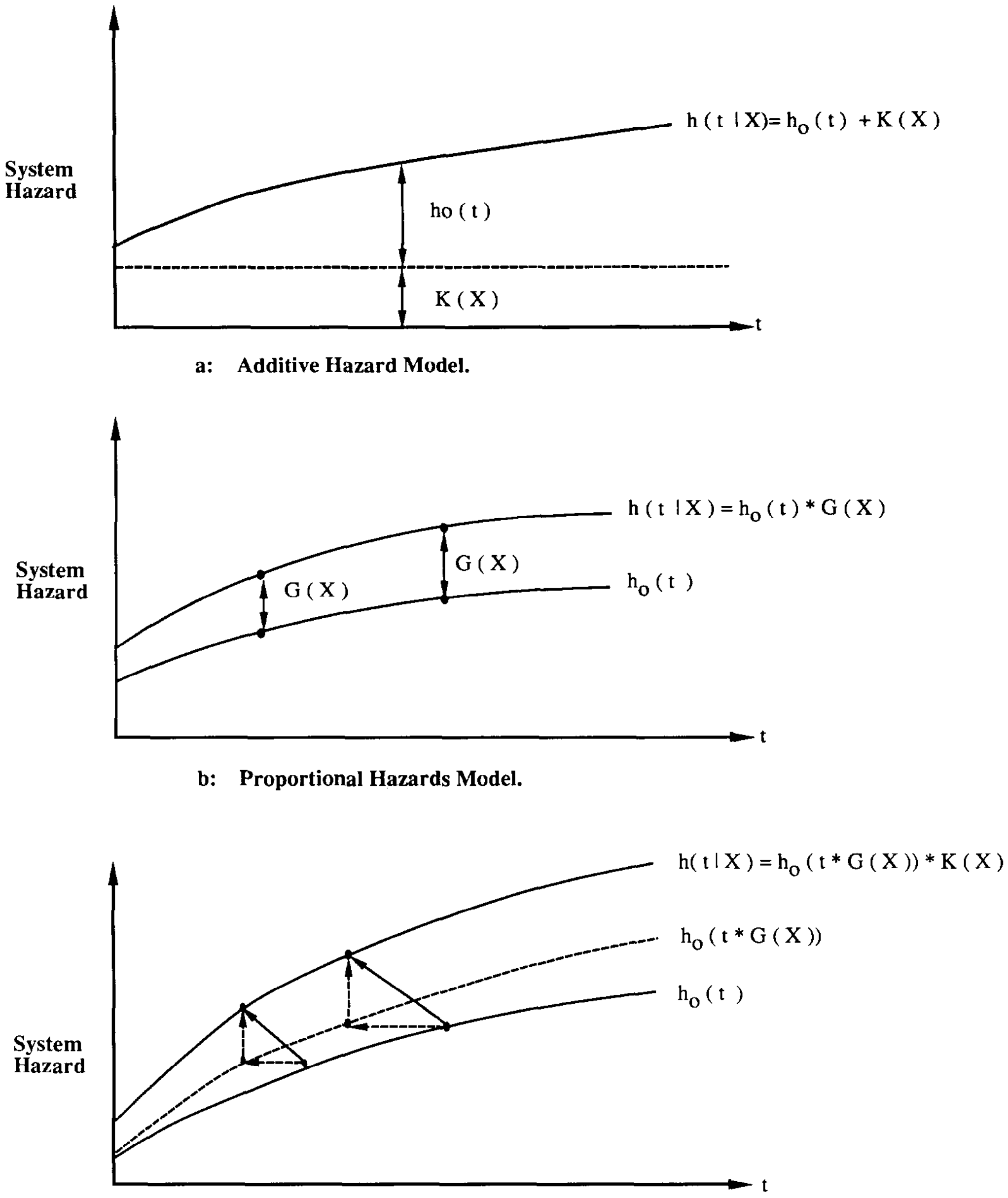
in which $h_0(t)$ represents the underlying hazard when $G(\mathbf{X}) = 1$. This model possesses the property that different individual trips have hazard functions that are proportional to one another, so it is also called the proportional hazard model. That is, the ratio $h(t|\mathbf{X}_1)/h(t|\mathbf{X}_2)$ of the hazard functions for two trips with constant regressor vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ does not vary with time. If the $G(\mathbf{X})$ in the proportional hazard model is further assumed to be a linear additive function of the risk components, then the proportional hazard model implies that all the specified risk components operate individually and their effect on the hazard function depends on $h_0(t)$. In essence, $G(\mathbf{X})$ is a multiplicative factor of the hazard function and must be a nonnegative function.

Cox (1972) suggested a particular form of $G(\mathbf{X})$ for the proportional hazard model:

$$h(t|\mathbf{X}) = h_0(t) * \exp(\mathbf{X}\boldsymbol{\beta}), \tag{7}$$

where $\mathbf{X}\boldsymbol{\beta} = x_1\beta_1 + x_2\beta_2 + \text{———} + x_p\beta_p$ and the β_is are unknown regression coefficients (see Fig. 1b). This model is natural and sufficiently flexible for many purposes. Since Exp($\mathbf{X}\boldsymbol{\beta}$) is always positive, $h(t|\mathbf{X})$ will be automatically nonnegative for all $\mathbf{X}$ and $\boldsymbol{\beta}$, if $h_0(t)$ is positive. This model is widely accepted in biomedical studies of the lifetimes of observed patients and is named the Cox model (e.g. Kalbfleisch and Prentice 1973, 1980; Holt 1978).

The Cox model possesses the specific characteristic that the increase in system hazard



a: **Additive Hazard Model.**

b: **Proportional Hazards Model.**

c: **Accelerated Failure Time Model.**

Fig. 1. Alternative hazard functions.

due to an increase in one risk component will depend on the level of all the other risk components. That is, when the risk component x_i increases Δx_i, the hazard function $h(t|\mathbf{X})$ will increase to $h_0(t) * \text{Exp}(\mathbf{X}\boldsymbol{\beta}) * \text{Exp}(\Delta x_i \beta_i)$. This implies that strong interdependence exists among the risk components. An increase in the level of one risk component will increase the system hazard (i.e. probability of an accident) but the increased hazard depends on the levels of the other risk components. This is a desirable property for accident occurrence models because this interdependence occurs frequently in accident etiology. For example, the hazard represented by heavy snow will vary depending on vehicle condition, driver condition, and many other risk components.

We can treat $h_0(t)$ as the combined hazard due to driver fatigue (from continuous driving) and the risk components that are not included in the model. Then $\text{Exp}(\mathbf{X}\boldsymbol{\beta})$ is the hazard adjusting factor for the risk components included in the model. The Cox model can thus be applied to simultaneously formulate the effects of driver's fatigue and other risk components on accident involvement.

A third type of hazard function is one in which the risk components alter the rate at which an individual (or vehicle trip) proceeds along the time axis. These models are called accelerated failure time models (see Fig. 1c). In eqns (6) and (7), it is clear that there is no direct relationship between $G(x)$ and t itself; the effect of regressor variables is to increase the hazard not to explicitly effect survival time. In accelerated failure time models, as the level of risk components increase, the system hazard function shifts horizontally to directly reflect accelerated failure time as well as vertically to reflect increased hazard. Accelerated failure time models are rarely used because of estimation difficulties arising from time-dependent covariates (Kalbfleisch and Prentice 1980).

Application to highway safety

Rather than model the disease processes of individuals, which typically progress over several years, we are seeking to model the occurrence of highway accidents for vehicle trips that occur over a time span of hours. One could formulate a safety model of individual drivers in such a way that the time to failure is the time between their accidents. Since this is likely to be on the order of months or years it is difficult (indeed impossible) to measure the effect of important risk components (such as roadway, environment, etc.) on accident occurrence.

An alternative view is to focus on individual trips which are more fully described by their attributes. A "failure" is thus defined as a trip during which an accident occurred, while a trip with no accident is considered censored (i.e. failed at some unknown time beyond the duration of the study). Since accidents are rare events this means that very few trips will result in failures. Fortunately, these failures are readily described in accident reports so that there is generally much information about failed trips. While there is generally limited information on successful or censored trips in the broad highway safety field, most large trucking companies retain extensive descriptions of these trips. By collecting data for a large number of accident and nonaccident trips one can obtain completely consistent descriptors of risk components for each of these trip types. By estimating the hazard function, $h(t|\mathbf{X})$, one can thus statistically determine the effect of risk components on accident occurrence by examining the sign, magnitude, and significance of regression parameters. The estimated hazard function thus represents the effect of the risk components on accident occurrence. This formulation achieves our objective of including accident (failure) and exposure (censored) data at a consistent level and should allow for clearer identification of the effect of risk components.

The next section formulates the estimation problem in general and then specifically for the Cox model, which seems most applicable to highway safety studies. Particular attention is paid to the effect on parameter estimation of using only accident data and using accident data with nonaccident observations. Extensions to this initial formulation are discussed in the section, Summary and Future Research Directions.

4. ESTIMATION

Consider a population of individual vehicle trips; for each trip we observe either the time until an accident (failure) or the time until the successful completion of a trip (censoring). The likelihood function for a set of censored data on n trips can be expressed as follows when the lifetime distribution of trips is considered to be a function of constant regression vector $\mathbf{X}_i$:

$$L[h_0(t), G(\mathbf{X})] = \prod_{i=1}^{n} [h_0(t_i) * G(\mathbf{X}_i) * S(t_i|\mathbf{X}_i)]^{\delta_i} * [S(t_i|\mathbf{X}_i)]^{1-\delta_i}, \qquad (8)$$

where t_i is the lifetime or censoring time for trip i and δ_i is the usual indicator variable taking on the value 1 if t_i represents an accident and 0 if t_i is censoring time.

In the parametric estimation approach, the distribution function of $h_0(t)$ and the multiplicative factor $G(\mathbf{X})$ are specified, and one estimates the parameters in both $h_0(t)$ and $G(\mathbf{X})$ simultaneously by maximizing the likelihood function eqn (8) for the observed data. Distribution-free methods are also available.

Suppose that a random sample of n individuals yields a sample with k distinct observed lifetimes and $n - k$ censoring times. The k observed lifetimes will be denoted by $t_1 < t_2 <$ ——— $< t_k$, and $R_i = R\,(t_i)$ is risk set at time t_i, that is, the set of individuals whose failure or censoring time is at least t_i. If we represent the hazard function by the Cox hazard model, i.e. eqn (7), then the likelihood function (8) of those n observed individuals can be written as:

$$\begin{aligned}
L[\boldsymbol{\beta}, h_0(t)] &= \prod_{i=1}^{n} \{h_0(t_i) * \exp(\mathbf{X_i}\boldsymbol{\beta}) * S(t_i|\mathbf{X_i})\}^{\delta_i} * \{S(t_i|\mathbf{X_i})\}^{1-\delta_i} \\
&= \prod_{i=1}^{n} \{h_0(t_i) * \exp(\mathbf{X_i}\boldsymbol{\beta}) * \exp[-H_0(t_i) * \exp(\mathbf{X_i}\boldsymbol{\beta})]\}^{\delta_i} \\
&\quad * \{\exp[-H_0(t_i) * \exp(\mathbf{X_i}\boldsymbol{\beta})]\}^{1-\delta_i} \\
&= \prod_{i\in D} \left\{\exp(\mathbf{X_i}\boldsymbol{\beta}) \Big/ \sum_{j\in R_i} \exp(\mathbf{X_j}\boldsymbol{\beta})\right\} * \prod_{i\in D} \left\{h_0(t_i) \sum_{j\in R_i} \exp(\mathbf{X}\boldsymbol{\beta}_\mathbf{j})\right\} \\
&\quad * \prod_{i=1}^{n} \{\exp[-H_0(t_i) * \exp(\mathbf{X_i}\boldsymbol{\beta})]\},
\end{aligned} \tag{9}$$

where D denotes the set of individuals who fail. It is assumed, in conjunction with the continuous model, that all lifetimes are distinct and that they are distinct from any censoring times. The function $H_0(t)$ is the cumulative nuisance hazard, defined similarly to the cumulative hazard of eqn (4).

Cox suggested a nonparametric approach to estimate the $\boldsymbol{\beta}$ in (7) using the product limit method (Kaplan and Meier 1958). He assumes that $h_0(t)$ is arbitrary, and no information can be contributed about $\boldsymbol{\beta}$ by time intervals in which no failures occur because the component $h_0(t)$ might conceivably be identically zero in such intervals (Cox 1972). The probability that an individual failure occurs at time t_i, conditional on the risk set $R(t_i)$, is a logistic form:

$$\mathrm{Exp}\{\mathbf{X_i}\boldsymbol{\beta}\} \Big/ \sum_{j\in R_i} \mathrm{Exp}\{\mathbf{X_j}\boldsymbol{\beta}\}. \tag{10}$$

Each failure contributes a factor of this nature and $\boldsymbol{\beta}$ can be estimated by maximizing the following conditional likelihood function:

$$L_1\,(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left\{\mathrm{Exp}(\mathbf{X_i}\boldsymbol{\beta}) \Big/ \sum_{j\in R_i} \mathrm{Exp}(\mathbf{X_j}\boldsymbol{\beta})\right\}. \tag{11}$$

Observe that (11) is the first term of the full maximum likelihood eqn (9). The Cox estimation approach, called a maximum partial likelihood estimation method, sacrifices the information contained in the last two terms of eqn (9).

Estimation of either the full likelihood (eqn (9)) or partial likelihood (eqn (11)) functions for safety studies requires careful consideration of the treatment of both accident and nonaccident data. When estimating the full or partial likelihood function, both accidents and nonaccidents enter the estimation directly. Accidents are such rare events, however, that simple random sampling is unlikely to result in sufficient accident trips for model estimation. Enriched sampling, in which accident trips are oversampled, can be used to provide accident data, but a large number of nonaccident trips will be required for model estimation. In a study of interstate truck accidents, it is estimated (Chang 1987) that at least 6,000 nonaccident trips are required to obtain sufficient statistical power.

Estimation of the partial likelihood, $L_1(\boldsymbol{\beta})$, appears to be simpler but is complicated by the fact that the denominator should contain trips with censoring times greater than t_i (i.e. both accident and nonaccident trips with longer lifetimes). Exclusion of these nonaccident trips may introduce unknown biases in parameter estimates and result in an inability to represent the survival process of the entire population.

Several advantages of the partial likelihood approach make it desirable for us to continue to explore its use in safety studies using accident trips only to estimate $L_1(\boldsymbol{\beta})$. First, eqn (9) is essentially an order statistics model that tries to determine why one vehicle trip has a shorter lifetime than the others. Considering the accident trip with lifetime t_i, all the accident trips with lifetimes greater than t_i are still surviving as nonaccident trips at time t_i. Ranking the lifetimes of accident trips may be thought of as being equivalent to taking the exposure of accident trips into account. It enriches the information derived from the accident data. Knowledge about how the risk components determine the lifetimes of accident trips should be reasonably extendable to the population as a whole. Second, if the number of accident trips is large enough, those accident trips could well reflect the accident occurrence of different patterns of trips in the population; the bias could then be very small. Obviously, eliminating the need for nonaccident (exposure) data collection can offer significant cost advantages.

The following sections describe an application of this method to a set of accident data. The analysis is undertaken to test the interpretability of the regression parameters represented by $L_1(\boldsymbol{\beta})$. After discussion of the data set and model results, there is an extended discussion of potential model applications in the safety field.

5. THE DATA SET

Accident reports from a major national carrier of less-than-truckload freight provide the primary data for model estimation. The accident data include time and location of accident, weather and lighting condition, highway type, type of collision, and severity of injury. In addition to characteristics of the accident, driver logs were used to obtain the number of hours on duty during the previous eight days and the number of hours off-duty prior to the accident trip. Both of these variables are indicators of driver fatigue. Table 1 summarizes the available data categorizing the risk components as constant for the entire trip, or situational along the trip.

Constant variables are easily included in the Cox model. The driving hour is recorded up to the time of the accident as the lifetime of the accident trip. The most difficult variables to include in the formulation are the situational variables. This is because these variables affect accident occurrence at the scene and possibly vary during the course of the trip. The model application does not include situational variables although the discussion of future research includes model extensions that may be able to capture the effect of situational conditions (e.g. roadway geometry) in the hazard function specification.

Data were collected for 1982 and 1983 accidents representing over 1,200 collisions. This trucking company operated a fleet of tractor semi-trailer combinations and double trailer combinations throughout the United States. Yearly vehicle miles exceeded 300 million using over 2,500 tractors and 10,000 trailers.

The accident data reflect all road accidents, excluding those during urban pick up and delivery. The company operates their trucks as "pony-express" service with one driver taking a tractor and trailer over one leg (generally less than 500 miles), then transferring the shipment to a new driver. The first driver rests (for at least eight hours) and then may take a return trip.

Table 2 contains the definitions of the variables used in the model application. The lifetime used in the models is recorded to the nearest one quarter hour. This information was derived directly from the drivers' daily logs. These data provide us with a description

Table 1. Types of risk components for motor carrier safety analysis

Type of Risk Component Factor	Constant	Situational
Vehicle	* Type of Truck * Weight of Cargo * Age of Tractor	
Driver	* Age * Experience * Accident Record * Off-duty Hours before starting the trip	
Roadway		* Type of Roadway * Width of Lane * Number of Lanes
Environment		* Weather * Lighting * Traffic Volume * Topography * Night or Daytime

of the ability of an individual truck (and driver) to "survive" in the highway system. By estimating the model as described in Section 4, we are able to identify the risk components that influence the lifetime of each truck trip.

6. MODEL APPLICATION

The purpose of these preliminary estimations is to determine if rational model parameters are obtained. A secondary objective is to see if the segmentation of accidents yields a different set of significant risk components. The pooled accident model is developed primarily for comparison with the segmented models. The results are obtained from the partial maximum likelihood estimation using eqn (11) with accident data only.

Table 2. Variable definitions

Variable name	Definition
WINTER	A dichotomous variable: 1 if accident occurred during December, January, or February; 0 otherwise
NIGHT	A dichotomous variable: 1 if the accident occurred from 9:00 P.M. to 6:00 A.M.; 0 otherwise
AGE	The age of the truck driver in years
EXPRNCE	Years of employment with the company that supplied the data
WEIGHT	Cargo weight in thousands of pounds
HRDUTY8	Number of hours on-duty and driving during last eight days including the day of the accident
HROFF	Number of consecutive hours off-duty prior to the start of the accident trip

Table 3. Model results: Pooled data

Variable	Coefficient	Standard error	t
WINTER	−.0785	.0592	−1.32
NIGHT	.0619	.0571	1.08
AGE	−.0009	.0034	−.27
EXPRNCE	−.0049	.0043	−1.16
WEIGHT	.0006	.0021	.30
HRDUTY8	−.0144	.0022	−6.43
HROFF	−.0022	.0007	−3.18

Log likelihood = −7783.7923; Global χ^2 = 50.03; *df* = 7; *p*-value = .000; Cases = 1,260.

Using all the predictors in Table 2, the Cox model is estimated on the entire data set. Two separate market segments are then explored: first, severity of collision characterized as property damage only and injury/fatality; second, type of collision characterized as single vehicle and multiple vehicle.

Pooled model

Pooled model results are shown in Table 3. Driver service hour variables are strongly significant and negative in sign. The interpretation is that drivers with higher driving hours in the previous eight days (HRDUTY8) have a reduced risk of an accident (thus a negative sign to the coefficient). Similarly, drivers with a large number of hours off-duty (HROFF) prior to the accident trip also have reduced accident risk.

The driver service hour variables appear to be capturing the influence of medium- and short-term driver scheduling policies. Frequently used drivers, those with high cumulative hours during the previous eight days, appear to have their driving skills well tuned and are surviving longer in the risk system than the less frequently scheduled drivers. Operating within this medium-term policy is the short-term scheduling of drivers for a particular trip: drivers with a high number of hours off-duty immediately prior to the trip survive longer. Trucking companies would appear to be able to reduce their accident risk if they had a limited number of frequently used drivers who were able to have more than the minimum eight hours off-duty prior to a trip.

These findings should not be interpreted as evidence to modify driver service hours regulations. Inspection of driver logs revealed that all accident-involved drivers were within U.S. Department of Transportation regulations for cumulative service hours (70 hours in 8 days). While there is some interrelationship between short-term hours off-duty and cumulative driving for the previous eight days, the statistical correlation is low (less than 0.5).

All other variables are statistically insignificant, particularly cargo weight and driver age. The effect of driver experience, winter, and night are also insignificant at any reasonable significance level. While these variables have limited association with overall accident risk, they are related to the risk of different types of collisions and their severity.

Accident severity

When data are segmented by severity of collision, some very significant differences occur in parameter estimates (see Table 4). For property damage only (PDO) accidents, cumulative work hours and hours off-duty have nearly the same sign and magnitude as in the pooled model. All other parameters are insignificant.

Marked changes occur for injury accidents (there was only one fatality in our data set). Hours off-duty is now insignificant while cumulative work hours is marginally significant though it retains a negative sign. Experience is strongly significant and negative, indicating that experienced drivers have lower risk of severe accidents than inexperienced drivers. Three other variables are of marginal significance. Night and age are positive indicating increased hazard for older drivers and driving at night. There is

Table 4. Model results: Accident severity

Property damage only

Variable	Coefficient	Standard error	*t*
WINTER	−.0699	.0666	−1.05
NIGHT	.0380	.0637	.60
AGE	−.0037	.0038	−.99
EXPRNCE	−.0011	.0047	−.23
WEIGHT	.0007	.0024	.28
HRDUTY8	−.0157	.0025	−6.28
HROFF	−.0022	.0007	−3.00

Log likelihood = −6036.4081; Global χ^2 = 44.59; *df* = 7; *p*-value = .000; Cases = 1,013.

Injury and fatality

Variable	Coefficient	Standard error	*t*
WINTER	−.2014	.1434	−1.40
NIGHT	.2184	.1397	1.56
AGE	.0128	.0086	1.49
EXPRNCE	−.0281	.0112	−2.52
WEIGHT	.0060	.0053	1.12
HRDUTY8	−.0099	.0054	−1.82
HROFF	−.0021	.0018	−1.15

Log likelihood = −972.3387; Global χ^2 = 16.79; *df* = 7; *p*-value = .000; Cases = 200.

Table 5. Model results: Collision type

Single-vehicle collisions

Variable	Coefficient	Standard error	*t*
WINTER	−.1517	.0863	−1.76
NIGHT	−.0181	.0867	−.21
AGE	−.0057	.0051	−1.11
EXPRNCE	.0026	.0064	.40
WEIGHT	.0017	.0030	.57
HRDUTY8	−.0125	.0033	−3.82
HROFF	−.0019	.0010	−1.98

Log likelihood = −3028.5001; Global χ^2 = 19.29; *df* = 7; *p*-value = .0073; Cases = 563.

Multiple-vehicle collisions

Variable	Coefficient	Standard error	*t*
WINTER	.0358	.0849	.42
NIGHT	.1544	.0773	2.00
AGE	.0036	.0046	.78
EXPRNCE	−.0106	.0058	−1.84
WEIGHT	.0003	.0031	.10
HRDUTY8	−.0170	.0031	−5.51
HROFF	−.0024	.0009	−2.50

Log likelihood = −3880.5141; Global χ^2 = 41.46; *df* = 7; *p*-value = .0000; Cases = 696.

a marginal reduction in risk during the winter and marginal increase in risk with high weight.

It is interesting that driver work hours are so strongly significant in PDO accidents. The risk of severe accidents is much more strongly related to driver experience than service hours. Environmental factors (night, winter) also are much more significant than for PDO though they are marginal statistically. These results tend to support the contention that a different set of risk factors may influence the pattern of accidents than influence accident occurrence.

Type of collision

Table 5 contains estimation results for segmentation of accidents by type of collision (i.e. single vehicle and multiple vehicle). As before, there is a different pattern of significant variables in each segment.

Driver service hours are very significant and negative for multiple-vehicle collisions. Experienced drivers have a reduced risk of accident occurrence but night conditions increase the risk of multiple-vehicle collisions. Age and winter are strongly insignificant, as is cargo weight. These results are similar to findings for injury accidents except for the significance of the nightime variable. It seems reasonable that the risk of a multiple-vehicle collision would increase at night when sight distances are reduced. There is also some evidence that the conspicuity of large trucks is poor during night time, increasing the likelihood of another vehicle running into a truck.

Driver service hours are significant for single-vehicle crashes; while the sign is the same as for multiple-vehicle crashes, the magnitude and significance are less. All other variables are insignificant except for winter, which is marginal statistically and negative in sign.

7. SUMMARY AND FUTURE RESEARCH DIRECTIONS

A model of accident occurrence has been formulated using principles of survival theory. The model predicts the probability of being involved in an accident at time t given that a vehicle has survived until that time. Several alternative functional forms are discussed including additive, proportional hazards and accelerated failure time models. Model estimation is discussed for the case in which both accident and nonaccident trips are included and for the case with only accident data (conditional analysis). As formulated, the model has the distinct advantage of being able to consider accident and exposure data at a disaggregate level in an entirely consistent analytic framework.

A conditional accident analysis is undertaken using truck accident data obtained from a major national carrier in the United States. Model results are interpretable and generally reasonable. Of particular interest is that segmenting accidents into several categories yields very different sets of significant parameters. Driver service hours seemed to most strongly effect accident risk: regularly scheduled drivers who take frequent trips are likely to have a reduced risk of an accident, particularly if they have a longer (greater than eight) number of hours off-duty just prior to a trip.

The modeling framework opens several areas of promising future research. One obvious area is the estimation of models that include nonaccident data. Disaggregate nonaccident data are available from trucking companies through records of their daily dispatches. Parameter estimates can then be compared for partial and full likelihood functions. Estimation of the complete likelihood function can be conducted using a flexible maximum likelihood estimation program such as CHOMP (Daganzo and Schoenfeld 1978). The major task is to consider the appropriate sampling method to obtain nonaccident trips. A similar methodology can also be tested with data for train or air accidents or any other accident process where the time to failure and censoring is readily known.

One particular modeling improvement would be the ability to incorporate some description of the risk of roadway elements encountered prior to the accident occurrence (or in the case of safe trips, the risk of elements encountered over the span of the successful trip). Geometric design conditions may vary greatly throughout a vehicle trip. Given the known analytic difficulties associated with model estimation using time-related risk factors, it may not be possible to describe the elements in detail as they occur during a trip. An alternative approach is to use an index, perhaps similar to Greenshields' Index, to describe the design characteristics of the route. Still another alternative is to use dummy variables to represent different classifications of routes (e.g. a category for 90% or more mileage on interstates). While this will not allow for the precise description of each geometric element, it may allow for an assessment of comparative accident risk on roadways of differing designs.

A third enhancement is the continued refinement of the conceptual structure with appropriate linkage to model specification. Refinements in model specification may allow us to better assess the contribution of individual risk factors to accident occurrence and to consider alternative model forms such as accelerated failure time models.

This application of survival theory appears to hold promise in the analysis of accident data and identification of factors associated with accident occurrence. Further empirical tests will help greatly in the assessment of the model's broad applicability.

Acknowledgements—The authors gratefully acknowledge the support provided by the research sponsors: Yellow Freight System; the United Parcel Service Foundation; and, the Gar Foundation of Roadway Express. The authors also express appreciation to an anonymous referree who provided useful observations on the formulation of the likelihood functions and the model specification.

REFERENCES

Aranda-Ordaz, F. An extension of the proportional-hazards model for grouped data. Biometrics 39:109–117; 1983.

Barlow, R.; Proschan, F. Statistical theory of reliability and life testing: Probability models. New York, Holt Rinehart and Winston; 1981.

Ceder, A.; Livneh, M. Relationship between road accidents and hourly traffic flow—I. Accid. Anal. Prev. 14:19–34; 1982.

Chang, H. A disaggregate survival model of motor carrier highway accident occurrence. Ph.D. dissertation. Evanston, IL: Northwestern University, Department of Civil Engineering; 1987.

Chapman, R. The concept of exposure. Accid. Anal. Prev. 5:95–110; 1973.

Chirachavala, T., et al. Severity of large-truck and combination-vehicle accidents in over-the-road service: A discrete multivariate analysis. Transpn. Res. Rec. 975:23–26; 1984.

Cleveland, D.; Kitamura, R. Macroscopic modelling of two-land rural roadside accidents. Transpn. Res. Rec. 681:53–63; 1978.

Cox, D. Regression models and life-tables. J. Royal Stat. Soc. 2:187–220; 1972.

Cox, D.; Oakes, D. Analysis of survival data. New York, NY: Chapman and Hall; 1984.

Daganzo, C.; Schoenfeld, L. CHOMP user's manual. Berkeley, CA: Department of Civil Engineering, University of California; 1978.

Foldvary, L. Road accident involvement per miles travelled. Accid. Anal. Prev. 11:75–99; 1979.

Hakkinen, A. Traffic accident and professional driver characteristics: A follow-up study. Accid. Anal. Prev. 11:7–18; 1979.

Hall, J.; Dickinson, L.; Truck speeds and accidents on interstate highways. Transpn. Res. Rec. 486:19–32; 1974.

Hammerslag, R.; Roos, J.; Kwakernaak, M. Analysis of accidents in traffic situations by means of multiproportional weighted Poisson model. Transpn. Res. Rec. 847:29–36; 1982.

Holt, J. Competing risk analysis with special reference to matched pair experiments. Biometrika 65:159–165; 1978.

Institute of Traffic Engineers. Transportation and traffic engineering handbook. Englewood Cliffs, NJ: Prentice-Hall, Inc.; 1976.

Ivey, D., et al. Predicting wet weather accidents. Accid. Anal. Prev. 13:83–99; 1981.

Jovanis, P.; Chang, H. Modeling the relationship of accidents to miles travelled. Transpn. Res. Rec. 1068:42–51; 1986.

Jovanis, P.; Delleur, J. Exposure-based analysis of motor vehicle accidents. Transpn. Res. Rec. 910:1–7; 1983.

Kalbfleisch, J.; Prentice, R. Marginal likelihood based on Cox's regression and life model. Biometrika 60:267–278; 1973.

Kalbfleisch, J.; Prentice, R. The statistical analysis of failure time data. New York: John Wiley & Sons; 1980.

Kaplan, E.; Meier, P. Nonparametric estimation from incomplete observation. J. Am. Stat. Assoc. 457–481; 1958.

Koornstra, M. Multivariate analysis of categorical data with applications to road safety research. Accid. Anal. Prev. 1:217–221; 1969.

Lavette, R. Development and application of railroad-highway accident prediction equation. Transpn. Res. Rec. 628:12–19; 1977.

Lawless, J.C. Statistical models and methods for lifetime data. New York: John Wiley & Sons; 1982.

Lieblein, J.; Zelen, M. Statistical investigation of the fatigue life of deep-groove ball bearing. J. Res. Nat. Bureau Standards 57:273–316; 1956.

Meyers, W. Comparison of truck and passenger-car accident rates on limited-access facilities. Transpn. Res. Rec. 808:48–60; 1981.

Montgomery, D.; Peck, E. Introduction to linear regression analysis. New York: John Wiley & Sons Inc.; 1982.

Oppe, S. The use of multiplicative models for analysis of road safety data. Accid. Anal. Prev. 11:101–115; 1979.

Prentice, R.; Kalblfeisch, J. Hazard rate models with covariates. Biometrics 35:25–39; 1979.

Ruijgrok, C.; Van Essen, P. The development and application of disaggregate Poisson model for trip generation. Paper presented at the 59th Annual Meeting of Transportation Research Board, Washington, DC; 1980.

Snyder, J.F. Environmental determinants of traffic accidents: An alternate model. Transpn. Res. Rec. 486:11–18; 1974.

Treat, J.R. et al. Tri-level study of the causes of traffic accidents: Final report. Springfield, VA: National Technical Information Service; 1977.