academicJournals

Vol. 13(4), pp. 27-43, 28 February, 2018
DOI: 10.5897/SRE2017.6549
Article Number: 101297A56272
ISSN 1992-2248

http://www.academicjournals.org/SRE

Scientific Research and Essays

Full Length Research Paper

Survival modeling of accident risks of vehicle drivers in Northern Region of Ghana

Alhassan Faisal[1*], Mamadou Lamine Diedhiou[2] and Katara Salifu[1]

[1]Department of Statistics, Faculty of Mathematical Sciences, University for Development, P. O. Box 24, Navrongo, Ghana.
[2]School of Business Studies, Wisconsin International University College, Ghana.

Recieved 30 October, 2017; Accepted 16 February, 2018

The primary objective of this article is to investigate how survival modelling can be used in traffic accident analysis to explain driver accident risk factors. Accident records of 398 drivers from 2007 to 2009 were obtained from Motor Traffic and Transport Department (MTTD), Ghana Police Service, Northern region. Cox proportional regression model was employed for the analysis using the SAS package. The conclusion was that Survival modelling promises to be a useful tool for road safety analysis and the most significant variables to the risks of accident were driver characteristics (age, gender, experience), their behaviour in traffic (speed, use of alcohol, use of safety belt), the nature of exposure (annual kilometreage, road surface condition) and vehicle characteristics (vehicle age, weight, tyres condition). Implementation of the findings of this study will enable policy makers put up better measures to reduce accident occurrence in the region in particular and the country as a whole.

Key words: Cox proportional model, accident risks, annual vehicle kilometreage, survival modelling, SAS package.

INTRODUCTION

Road transport is a predominant means of commuting in Ghana accounting for high passenger travels and carting of goods in the country. Road transportation facilitates the movement of people, goods and services in all sectors of the economy, including tourism, mining, trade, health, education and agriculture, among others. Similarly, road crashes has also become a major national issue receiving front-page coverage in the press and National TV news on a regular basis.

Drivers are faced with risky situations and potential accidents every time they are on the road. Counter measures are taken by society to prevent accidents or moderate their consequences (Hakkanen and Summala, 2001). Accidents happen when road users cannot adapt their actions to the varying demands of the traffic environment. Consequently, the risk of accident can be lowered by improving road users' performance in traffic or by reducing system demands on road users (Elvik, 1996). Many factors affect driver accident involvement as found in literature. Factors may be considered accident causes if they either increase or decrease the probability of accident occurence. Therefore, to prevent accident, one must know which of the numerous traffic risk factors have a real strong influence on the number and probability of accidents. At any given time, driver accident risk is affected by personal risk factors, vehicle risk factors,

*Corresponding author. E-mail: alhassanfsl83@gmail.com.

environmental factors, and other risks created by other drivers and traffic (Elvik, 1996).

Past researches (Chieng-Meng et al., 2016; Elvik et al., 2004; Sullman et al., 2002; Häkkänen and Summala, 2000; Dagan et al., 2006; Taylor and Dorn, 2006; Rodríguez et al., 2003; Jovanis et al., 1991) have determined that over speeding, age, experience, sleep quality, driving mileage, vehicle weight, limited stopping distances, substantial traffic volume,the habits of the drivers and time of day have some relationship with the occurrence of road accidents.

Predictive accident models have been developed by various authors in the world. For example, Calliendo et al. (2013) studied crash prediction model for road tunnels. They used bivariate negative binomial regression, jointly applied to non-severe and sever crashes, to model the frequency of accident occurence. The regression parameters were estimated using the maximum likelihood method.

Oppe (1989) used multiple linear regression models, where the dependent variable (either number of accidents or accident rate) is a function of a series of independent variables such as speed or traffic volume. These models assumed the occurrence of accidents to be normally distributed and therefore, they lack the distributional property necessary to describe adequately the random and discrete vehicle accident events on the road and hence are inappropriate for making probabilistic statements about accident occurrence.

Saccomanno and Buyco (1988) and Blower et al. (1993) used a Poisson loglinear regression model to explain variations in accident rates. This regression model is especially suitable for handling data with large numbers of zero counts and therefore, cannot be appropriate for road accident counts, since it fails to account for extra-Poisson variation (the value of the variation could exceed the value of the mean) in the observed accidents counts.

To solve this problem of extra-Poisson variation, several authors such as Miaou (1994) developed two types of negative binomial models, one using a maximum likelihood method and one using a method of moments. The maximum likelihood model was found to be more reliable than the Poisson regression model in predicting accidents where overdispersion is present.

In 1949, R. J. Smeed also developed a log-linear regression model and he found an inverse relationship between the traffic risk (fatality per motor vehicle) and the level of motorisation (number of vehicles per inhabitant). This means that with annually increasing traffic volume, fatalities per vehicle decrease. Smeed concluded that fatalities (F) in any country in a given year are related to the number of registered vehicles (V) and population P) of that country by the following equation;

$$\frac{F}{V} = \alpha\left(\frac{V}{P}\right)^{-\beta} \quad (1)$$

where F = number of fatalities in road accidents in the country, V = number of vehicles in the country, P = population of the country, α = 0.003 and β = 2/3. This formula became popular and has been used in many studies. It is often called as Smeed's formula.

It was generally observed in the literature that the development of accident prediction models have largely been focused on parametric modeling (Collett, 2003), where the functional form of the model is completely specified. These models will be appropriate if only we are sure of the model was correctly specified. However, if we are not completely certain, as is typically the case, then the semiparametric survival modeling approach proposed by Cox (1972) will be most appropriate. It is a “robust” model in the sense that it provides results that closely approximates the true parametric model, and therefore, the user does not need to worry about whether a wrong parametric model is chosen.

The objectives of this study therefore, is to investigate how the principles of survival modeling can be used in modeling accidents occurrence and to identify the factors that influence accident risks of drivers. The survival modeling (Klembaum, 1996) approach has not been widely adopted by researchers in the area of accident data anlysis. Survival modeling is commonly applied in medicine to the study of serious diseases and treatment methods. This study will assess the development needs of survival models in the area of traffic accident analysis and the findings can serve as a basis for health care professionals and policy makers to create preventive measures for traffic accidents.

METHODS

Modeling approach

In order to achieve the set objectives of this research, we formulated the following two specific research questions. Can driver involvement in road traffic accident be examined with survival models? Do driver characteristics and behaviour (such as driver’s sex, age, experience, use of belt, use of alcohol, route failiarity, speeding), vehicle characteristics (such as vehicle’s age, weight, tyres thread, ownership, annual kilometrage) and traffic environment characteristics (such as accident scene, road surface condition, other traffic demands) contribute to accident risk? These questions will be answered using the developed survival models and the estimated parameters.

The general procedures of modeling the accident data is summarized in the following seven steps;

Step 1: MTTD Data collection and processing.

Step 2: Study the varibales in the data and their categorization/Codes.

Step 3: Preliminary analysis of the data is performed using the Kaplan-Meier estimate of survival curves and log-rank test (Kaplan and Meier, 1958). This univariate anlysis was perforemed to ascertain the significance of the variables under study and to use it as a basis for inclusion or otherwise of the the covariates in the final model.

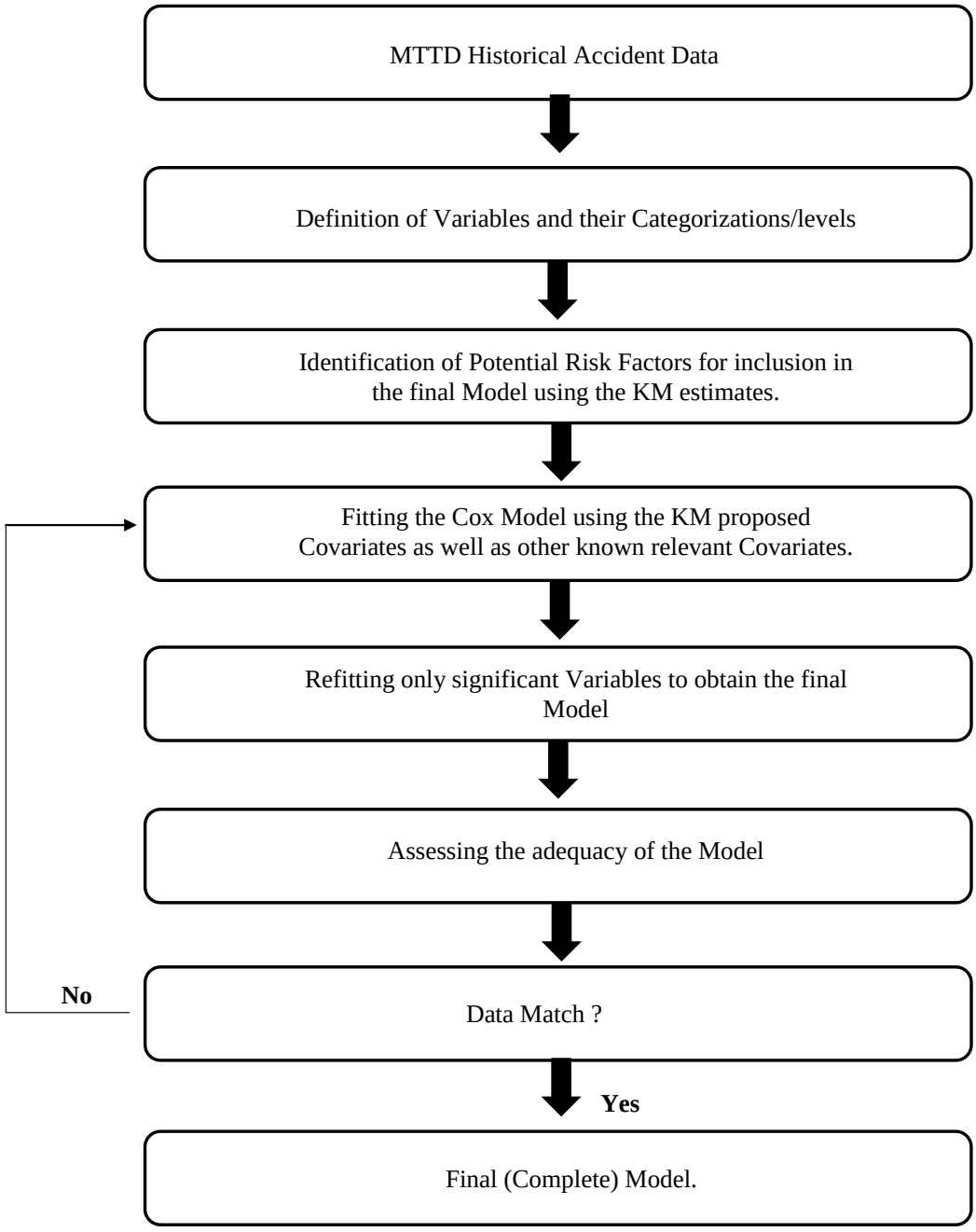

Figure 1. Flowchart of the modeling procedure.

Step 4: Fit the Cox model using the signifcant covariates suggested in step 3 and including other relevant variables using the Maximum Partial Likelihood Estimate (Cox, 1975).

Step 5: Refit the Model with only significant variables obtained in step 4 to obtain the final model.

Step 6: The final model is then evaluated to ascertain the goodness of the fit of the model.

Step 7: If the model fits the data well, the final model will be considered the complete model for the accident data, otherwise, we return to step 4 to refit the model using transformed values of the variables.

The flowchart of the modeling procedure is illustrated in the following Figure 1.

Source of data

Three years data containing detailed information on 398 accidents that involved drivers for the period of 2007 to 2009 were taken from Motor Traffic and Transport Department (MTTD) of the Ghana

Police Service, Northern Regional Office, Tamale.

The subsequent definitions and sentences in the *"Methods"* are mainly summary based on textbooks by Kalbfleisch and Prentice (2002), Klein and Moeschberger (1997), Allison (1995), Klembaum (1996) and Lawless (1982).

Survival time distribution

Survival analysis can simply be defined as time-to-event analysis (Klembaum, 1996); for example, time to die from disease say cancer. Survival data can be generated by observing a set of individuals at some well-defined point in time, and are followed for some substantial period of time, recording the times at which the events of interest occur and possibly some covariates associated with the individual that the risk of the event possibly depends upon.

But an important issue in survival research is how to deal with individuals whose survival cannot be followed during the entire research period. Such indivividuals are called censored individuals. There are generally three reasons why censoring may occur; a person does not experience the event before the study ends**,** a person is lost to follow-up during the study period, a person withdraws from the study because of some other reasons but not the event of interest. Censoring can happen in the following three ways;

Type I: the duration of the study is fixed to a chosen period. The individuals are monitored from a set starting point and individuals who are lost to the monitoring, or are withdrawn from the study or do not experience the event at the end of the study period, are censored observations.

Type II: The length of the monitoring period depends on the desired number or proportion of uncensored observations. The length of the study period is the same as the survival of the individual with the longest life span. Individuals, who are removed from the study for various reasons or survive less than the monitoring period, are censored observations.

Type III: The duration of monitoring is fixed. However, individuals may enter the study at different starting points. Censored observations are the ones whose survival period continues after the overall monitoring period has ended.

Survival studies can be divided into the following two groups:

i) Monitoring censored to the right: the investigation has begun at a certain selected moment when the individuals entering the examination are exposed to the phenomenon under investigation, e.g. a medicine or treatment, the investigation is continued from that moment on for a certain length of time.

ii) Monitoring censored to the left: the investigation has begun at a certain selected moment, but includes indivividuals whose exposure to the phenomenon under investigation has begun before the examination period started (as in the present study).

The Greek letter delta (δ) denote a {0, 1} random variable indicating either failure or censorship. That is, δ = 1 for failure if the event occurs during the study period, or *δ* = 0 if the survival time is censored by the end of the study period.

The survival time *T* , can be assumed to be following either a certain distribution or by direct observation based on the actual data. The most commonly used survival distributions are the negative exponential distribution, the Weibull distribution, the Gumbel distribution, the Logarithmic normal distribution or their combinations. The type of distribution that is best at describing the survival distribution is mainly dependent on the data.

If *T* represents a continous survival time, then its distribution is characterized by three functions; the survival function, *S*(*t*), which gives the probability that a person survives longer than some specified time *t,* that is, $S(t) = P(T > t)$. The probability density function, $f(t)$, which gives the probability of an individual experiencing the event of interest at exactly time *t*, where $f(t) = F'(t) = -S'(t)$, where $F(t) = P(T \le t)$, is the cumulative distribution function, which gives the probability of an individual not surviving beyond *t.*

The hazard function $h(t)$, on the other hand, gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time *t.* The survival function, *S*(*t*), is most useful for comparing survival progress of two or more groups whilst the hazard function, $h(t)$, gives a more useful description of the risk of failure at any point in time. Since *T* denotes time, it can be any number equal to or greater than zero and a *t* denotes any specific value of interest for the random variable *T*. The following relationship exist between these functions:

$$\left.\begin{array}{l} h(t) = \dfrac{f(t)}{S(t)} = \dfrac{-d\log S(t)}{dt}, \text{ but} \\ S(t) = 1 - P(T \le t) = 1 - \mathrm{F}(t). \\ \text{Also, } S(t) = \ell^{-\int_0^t h(t)dt} = exp[-\int_0^t h(t)dt] \end{array}\right\} \quad (2)$$

where $\int_0^t h(t)dt = H(t)$ is the cumulative hazard function.

This means that

$$S(t) = exp[-H(t)] \quad (3)$$

If one of these functions is known, the other two can be determined.

The Cox Proportional Hazards model and its characteristics

The survival model type that will be used is the Cox Proportional regression model (Cox, 1972) to examine driver's accident risks and their dependence on characteristics connected with drivers and vehicles, as well as the prevailing road way conditions.

The Cox PH model is usually written in terms of the hazard model formula

$$h(t, X) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) = h_0(t)exp(\beta' X) \quad (4)$$

where *h(t,X)* is the hazard function, that is, hazard at time *t* for an individual with a given specification of a set of explanatory variables, X which are assumed to be time-independent, $h_0(t)$ is the base level of the hazard function; which is the hazard function for an individual, prior to considering any of the *X*'s (it represents the nonparametric part of the model) and can be thought of as the intercept in multiple regression. The $exp(\beta' X)$ is linear function formed by the variables and their parameters representing the parametric part of the model. It is this property that makes the Cox regression model a semiparametric model. The measure of the effect is called hazard ratio. The hazard ratio (HR) of two individuals with different covariates X and X^* is

$$HR = \frac{h_0(t)exp(\widehat{\beta}' X)}{h_0(t)exp(\widehat{\beta}' X^*)} = exp(\sum \widehat{\beta}'(X - X^*)) \quad (5)$$

This hazard ratio depends only on the predictor variables and not on time hence it is time-independent, which is also why the Cox regression is called the proportional hazard model. For indicator or (dummy) variables with values 1 and 0, one can interpret the hazard/risk ratio (ℓ^{β}) as the ratio of the estimated hazard for those with a value of 1 to the estimated hazard for those with a value of zero (controlling for other covariates). However, for quantitative covariates, a more helpful statistic is obtained by subtracting 1.0 from the hazard ratio and multiplying by 100. This gives the estimated percentage change in the hazard for each one unit increase in the covariate, holding other covariates constant (Allison, 1995).

The corresponding survival function for the Cox Proportional Hazard regression model is related as

$$S\left(t, X\right) = S_0(t)^{exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} = S_0(t)^{exp(\sum_{i=1}^{k} \beta_i x_i)} \quad (6)$$

The "partial" likelihood function is usually used instead of the "complete" likelihood function to estimate the parameters of the Cox model, because the likelihood formula considers probabilities only for those individuals who experienced the event, and does not consider probabilities for those individuals that are censored. The partial likelihood function for Cox PH model (Cox, 1975) is

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp(\beta' X_i(t_i))}{\sum_{k \in R(t_i)} exp(\beta' X_k(t_i))} \right]^{\delta_i} \quad (7)$$

where $R(t_i)$ is the risk set at time t_i. $X_i(t_i)$ is the vector of covariate values for individual i who dies at t_i, n is the observed survival time for n individuals and δ_i is the event indicator, which is zero if the i^{th} survival time is censored and unity otherwise. The partial likelihood is valid when there are no ties in the dataset. That means there are no two individuals who have the same event time.

Proportional Hazards (PH) assumption checking

The goal of statistical model development is to obtain the model which best describes the data. That is to say, the fitted model must provide an adequate summary of the data upon which it is based. Therefore, a complete and thorough examination of the model's fit and adherence to the model's assumption is of great importance and concern. The Cox PH model assumes that the hazard of one individual is proportional to the hazard of any other individual, where the proportionality constant is independent of time. This means that the ratio of the risk of accident of two drivers is the same no matter how long they have been driving. This requires that covariates not be time-dependent. If any of the covariates varies with time, the Proportional hazards assumption is violated.

There are several methods for verifying that a model satisfies the assumption of proportionality; they are the Graphical method, the method of adding time-dependent covariates in the Cox model, and tests based on the Schoenfeld residuals (Schoenfeld, 1982). In this study, the method of adding time-dependent covariates (Crowley and Hu, 1977) in the Cox model was employed. This can be done by including a time-covariate interaction terms in the model and testing if the coefficient for interaction is significantly different from zero. If a time-dependent covariate is significant, this indicates a violation of the proportionality assumption for that specific predictor. In this analysis, the interactions with log (time) was used because this is the most common function of time used in time-dependent covariates but any function of time could be used.

Cox Proportional Hazards model diagnostics

Several methods can be used to check the adequacy of a Cox PH model. We have the Cox-Snell Residual method (Cox and Snell, 1968), the Deviance residual method (Thernaeau et al., 1990), the Schoenfeld residual method (Schoenfeld, 1982) and Diagnostic for influential observations (Cain and Lange, 1984). In this study, the diagnostic for influential observations (Cain and Lange, 1984) was employed. This method is used to identify which if any observations, exert an undue influence on the estimates of the parameters and for that matter the fit of the model. According to Cain and Lange, an observation is said to be influential if removing the observation substantially changes the estimate of the coefficients.

The delta-beta (DfBeta) statistics is what is considered in this research and it tells one how much each coefficient will change by removal of a single observation. Therefore, we can check whether there are influential observations for any particular explanatory variable. The signs of the DfBeta statistics are the reverse of what one might expect – a negative sign means that the coefficient increases when the observation is removed.

Fitting Proportional Hazards Model for the MTTD data

The accident data contains, by definition, only drivers involved in accidents; this means that they have all experienced the event of interest. In order to analyse the data using survival analysis, it was assumed that all the drivers entered the study at the first day of the year and the survival, or accident time, was defined as the number of days counted from 1st January of the year to the day the accident occured. Drivers that were involved in accidents within the first 244 days (that is, first eight months) of each year were considered as "uncensored drivers" and those involved after the 244 days were considered as "censored drivers". Since this approach enabled a distinction of which observations should be classified as censored or uncensored, then the MTTD accident data can now be analyzed using survival approach. The theoretical display of this description is depicted in the Figure 2 (Allison, 1995).

The horizontal axis represents survival time (in days). Each of the horizontal lines labeled A through F represents a single driver. An x indicates that the accident occurred at that point in time. The vertical line at 244 is a point at which we stop following the driver. Any accident that occurred at 244 days or earlier were considered uncensored and those that occurred after 244 days are censored drivers. Therefore, drivers A, C, D and E have uncensored accident times while drivers B and C have censored accident times.

RESULTS

Preliminary analysis

In any data analysis, it is always a great idea to do some univariate analysis before proceeding to more complicated models. In survival analysis, it is highly recommended to look at the Kaplan-Meier (KM) curves and log-rank tests (Kaplan and Meier, 1958) for all the categorical predictors. The KM curves will provide insight into the shape of the survival functions for the categories

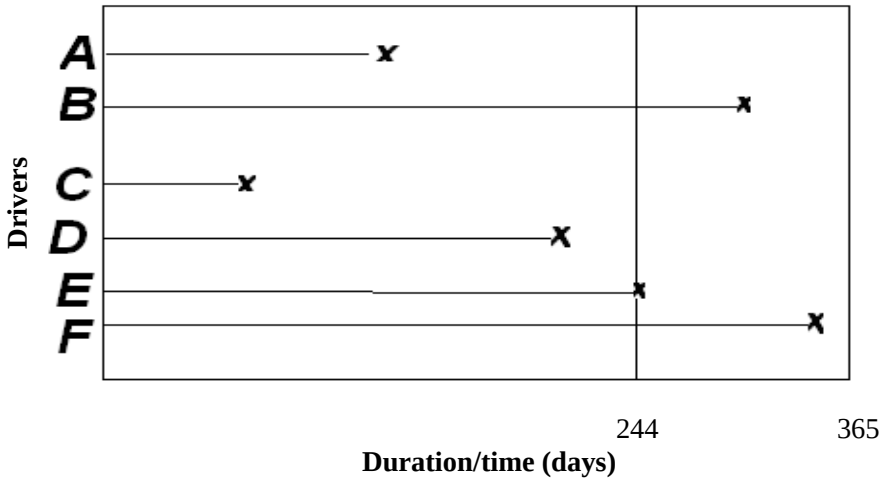

Figure 2. Theoretical display of survival time in the accident data.

of the variables to determine pictorially the survival experience of its categories as well as their proportionality (that is, the survival functions are approximately parallel). The KM test of equality across strata (categories), called the log-rank test, was employed in the preliminary analysis to explore whether or not to include the predictor in the final model. A variable was considered for inclusion into the final model if it is significant, that is, if the log-rank test has a p-value of 0.05 or less. If a predictor is not significant in a univariate analysis, it is highly unlikely that it will contribute anything to the model when it included in the final with other predictors.

Detailed information on the list of variables, the variables'categorizations/codes, reference categories of the variables, total frequencies of accidents under each variable, the number uncensored (that is, the number of drivers involved in accident within the first 8 months) and log-rank tests of equality across strata (categories) for the MTTD data are provided in the following Table 1.

The Kaplan-Meier survival curves (Figures 3 to 9) for only significant variables indicated from the above Table 1 in order to compare the survival experience of the different levels of the variables are as follows.

Fitted Cox Proportional Hazard Models

The Cox regression model (Cox, 1975) was first fitted to obtain model 1A displayed in Table 2 using the most important variables from the point of view of Kaplan-Meier test as well as other interesting variables. The significant variables in model 1A re-runned to arrive at the final model 1B.

This final model 1B was then evaluated by checking for the proportionality assumption (Table 5) and influential observations (Table 7) in the dataset.

DISCUSSION

Interpretation of the above outputs

Table 1 presents detailed information on Kaplan-Meier estimates for all the variables under study, the categories of each variable and their associated accident frequencies and reference levels. It can be seen that the variables that significantly contribute to accident time or probability according to the Kaplan-Meier estimate include driver age, sex, use of safety belt, use of alcohol, speed, age of driving license and age of vehicle since the log-rank test of equality across levels for each of them resulted in a p-value less than 0.05. The Kaplan-Meier survival curves that compared the survival experience of the levels of these significant variables are indicated in Figures 3 through 9.

These significant variables from the point of view of Kaplan-Meier and these relevant variables: annual vehicle kilometreage, tyres condition, weight of vehicle and route familiarity, were used to estimate Cox regression model 1A (Table 2). The model provides the coefficient estimates and their associated standard errors and significance. It is noted that there is no intercept estimate – a characteristic feature of partial likelihood estimation (Allison, 1995). The hazard ratios along with their 95% confidence interval are also shown in the last three columns.

The variable driver_agp had three levels and so the estimated coefficients for the last two levels as indicated in model 1A are supposed to be compared with that of the omitted first level which is the reference level. The same explanation can be given to the variable annukil as indicated in the same model. However, a more useful is the global test reported for both variables at the bottom part of the model 1A (Table 3). It can be seen that both variables driver_agp and annukil are significant at the

Table 1. List of variables and their categories, codes, reference category and log-rank tests for the MTTD data.

Variable code	Variable Name	Categories	Reference category	Total	Uncensored	p-value
driver_agp	Age of driver	1. ≤25 years	≤ 25 years	93	28	<0.0001
		2. 26 – 50 years		272	179	
		3. >50 years		33	14	
Sex	Sex of driver	1. Male	Male	377	203	0.0049
		2. Female		21	18	
Alcohol	Use of alcohol	0. No	No	160	50	<0.0001
		1. Yes		230	168	
usebelt	Use of safety belt	0. No	No	233	162	<0.0001
		1. Yes		109	29	
annukil	Vehicle annual kilometereage	1. < 5000 km/a	<5000 km/a	88	53	0.4016
		2. 5000 - 14,999 km/a		237	128	
		3. ≥15000 km/a		64	33	
roadsec	Road section/scene of accident	1.Links	Junction	304	159	0.0807
		2. Junction		72	47	
speedveh	Estimated speed of vehicle at time of accident	1. ≤ 80 km/h	>80 km/h	143	37	<0.0001
		2. >80 km/h		241	178	
wghtveh	Vehicle weight	1. ≤1,000 kg	>1,000 kg	212	115	0.4483
		2. >1,000 kg		174	99	
tyrescon	Tyres tread depth	0. ≤ 4 mm	>4 mm	278	152	0.3897
		1. >4 mm		111	68	
Agelic	Age of driving license	1. ≥5 years	≥5 years	295	180	<0.0001
		2. <5 years		80	30	
Rutfam	Route familiarity	0. Seldom pass scene	Seldom pass scene	61	33	0.8079
		1. Frequent		292	163	
roadsurf	Road surface condition	1. Dry	Dry	324	177	0.5470
		0. Wet		21	10	
vehown	Vehicle ownership	1. Own	Own	297	154	0.1280
		0. Not own		72	45	
ageveh	Vehicle age	1. ≤ 10 years	≤ 10 years	142	58	<0.0001

0.05 significance level. A variable is considered significant if its p-value is less than or equal to 0.05. This means that the highly significant variables in model 1A include use of safety belt, use of alcohol, speed, tyres condition, age of driver and annual vehicle kilometreage. The variables that were not significant and for that matter not qualified for inclusion in the final model include sex, vehicle age, age of driving license, weight of vehicle and route familiarity. However, the predictor sex is proven to be a very important variable to have in the final model and therefore it was added to the significant variables and the model was re-fitted to obtain the final model 1B (Table 4).

Hazard ratios/Relative risks to drivers

Driver age

From these models (Tables 2 and 4), it can be seen that drivers' age proved to be a major significant accident risk factor. In the model 1A which indicated the individual contribution of each category of the age groups, it can be seen that drivers aged between 26 to 50 years had 1.854 times (with 95% confidence interval: 1.136 to 3.027) riskier than drivers aged ≤ 25 and those who were older than 50 years had 2.170 times greater than those in their early 20s.

This conclusion is confirmed in Figure 3 which shows the survival function for each age group. It can be seen that the survival function for those drivers aged between 26 and 50 had higher share of accidents, followed by those above 50 years and up to or less than 25. In general, the pattern of one survivorship function lying above another means that the group defined by the upper curved live longer or had a more favourable survival experience than the group defined by the lower curve. This may reflect lack of experience (and perhaps riskier driving style) of young drivers, as for the old drivers it can

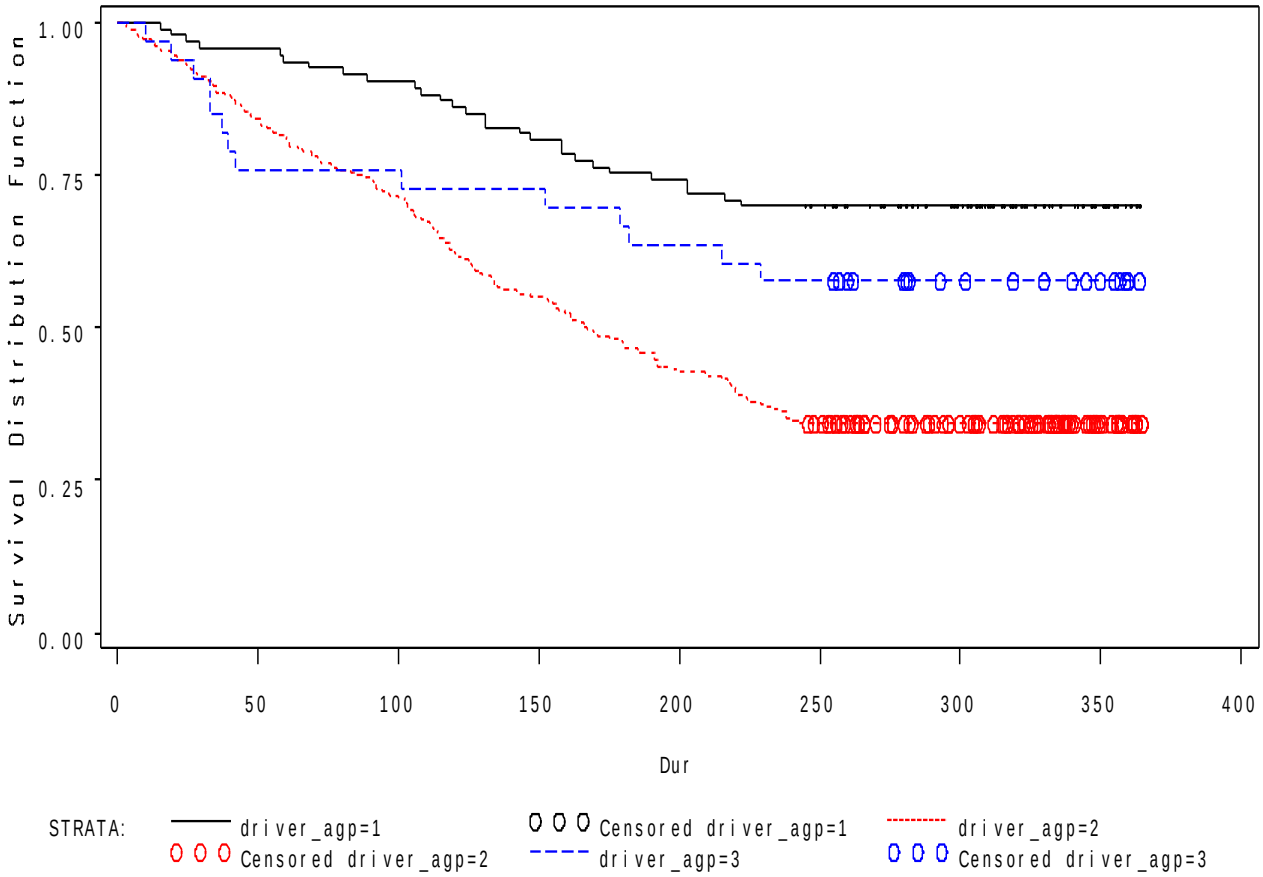

Figure 3. Survival distribution of the age groups of drivers.

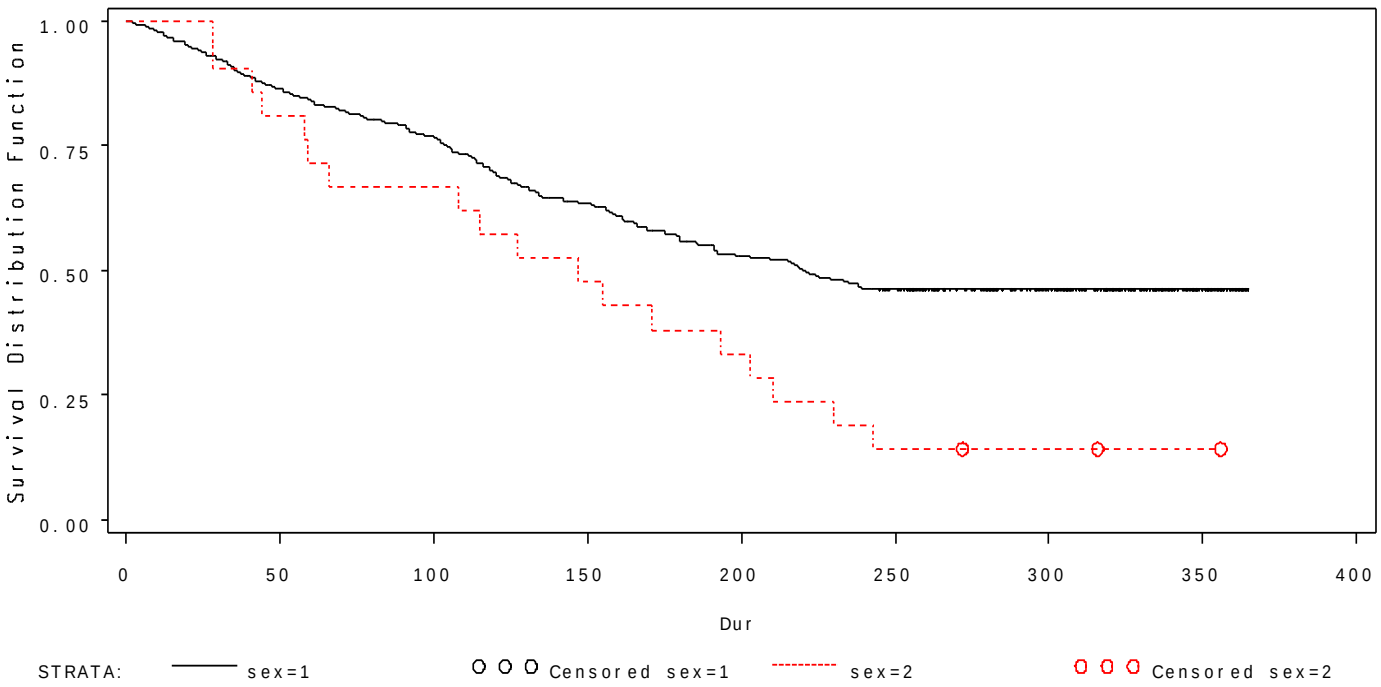

Figure 4. Survival distribution of the sex of drivers.

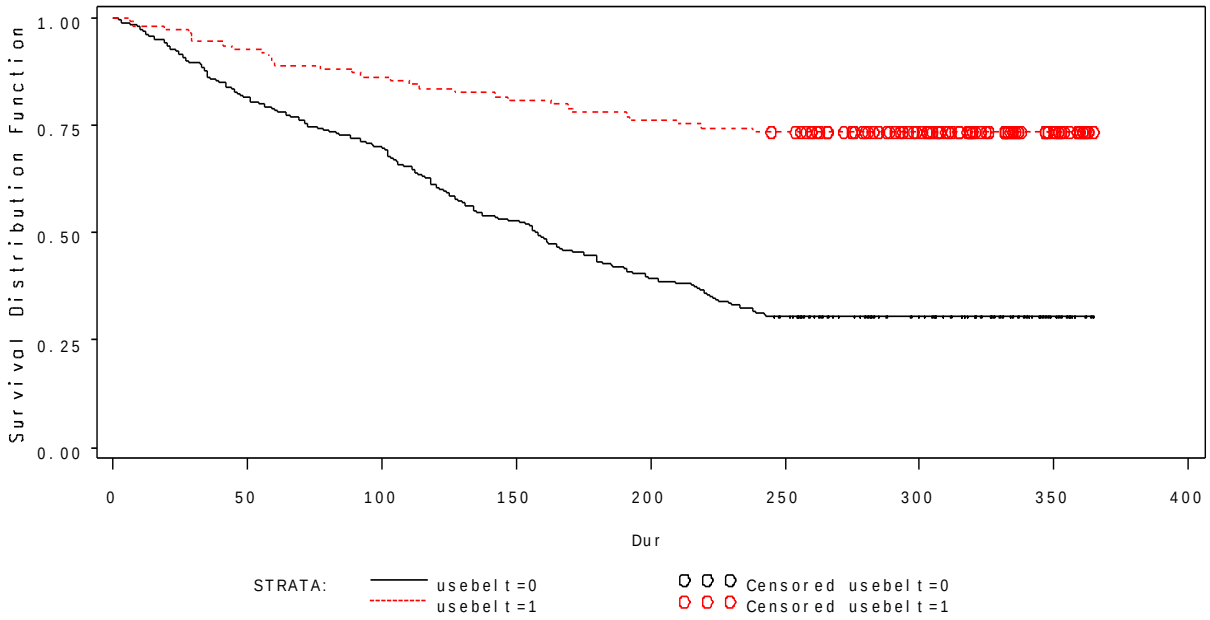

Figure 5. Survival distribution of the use of seat belts status of drivers.

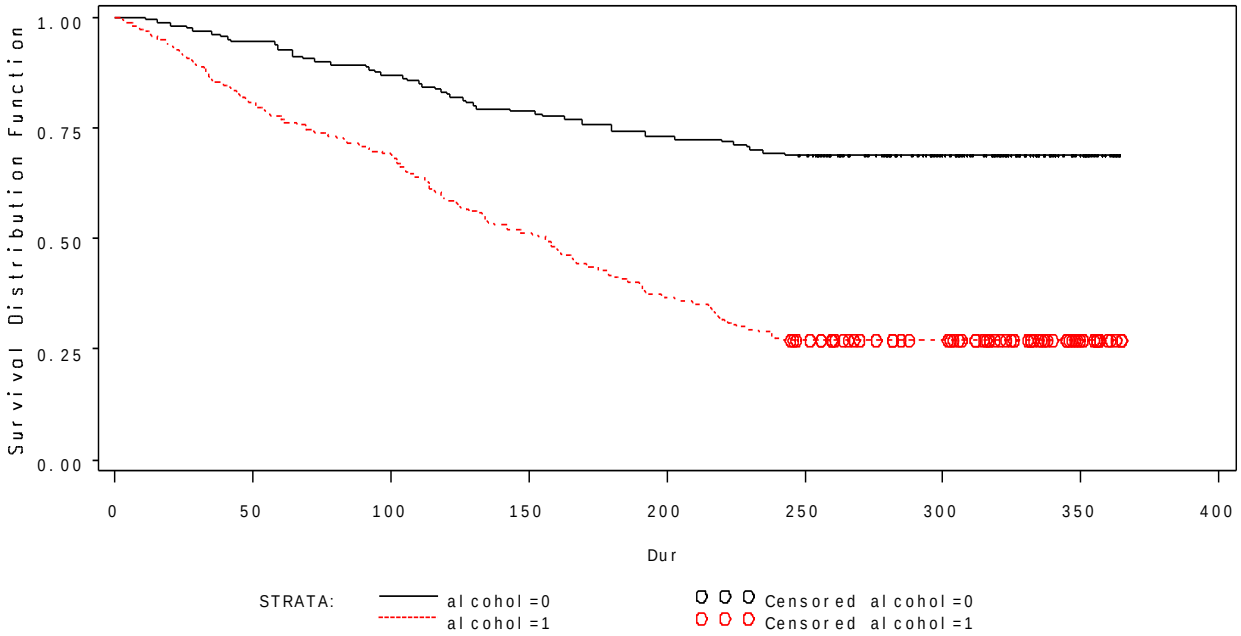

Figure 6. Survival distribution of the alcohol status of drivers.

be attributed to reduce capabilities on their part. Traffic conditions set greater demands on all drivers; the young and very old may have a harder time meeting the greater demands. Also, the final model 1B gave the overall contribution of driver age (p=0.0146) with a hazard ratio of 1.473 which means that for any one year increase in the age of the driver, The hazard of accident increases by 47.3%.

Driver sex

According to the Kaplan-Meier estimate, sex had a strong effect on accidents time. However, the effect of sex had a

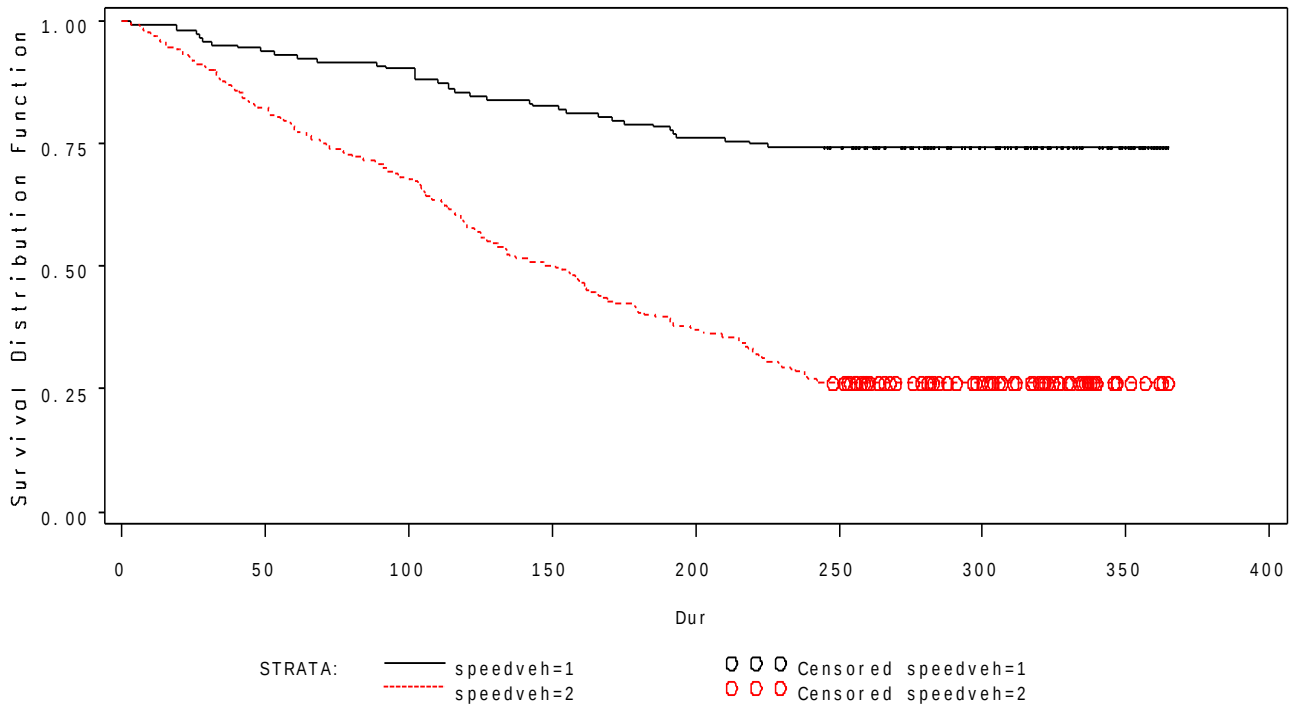

Figure 7. Survival distribution of the speed of drivers

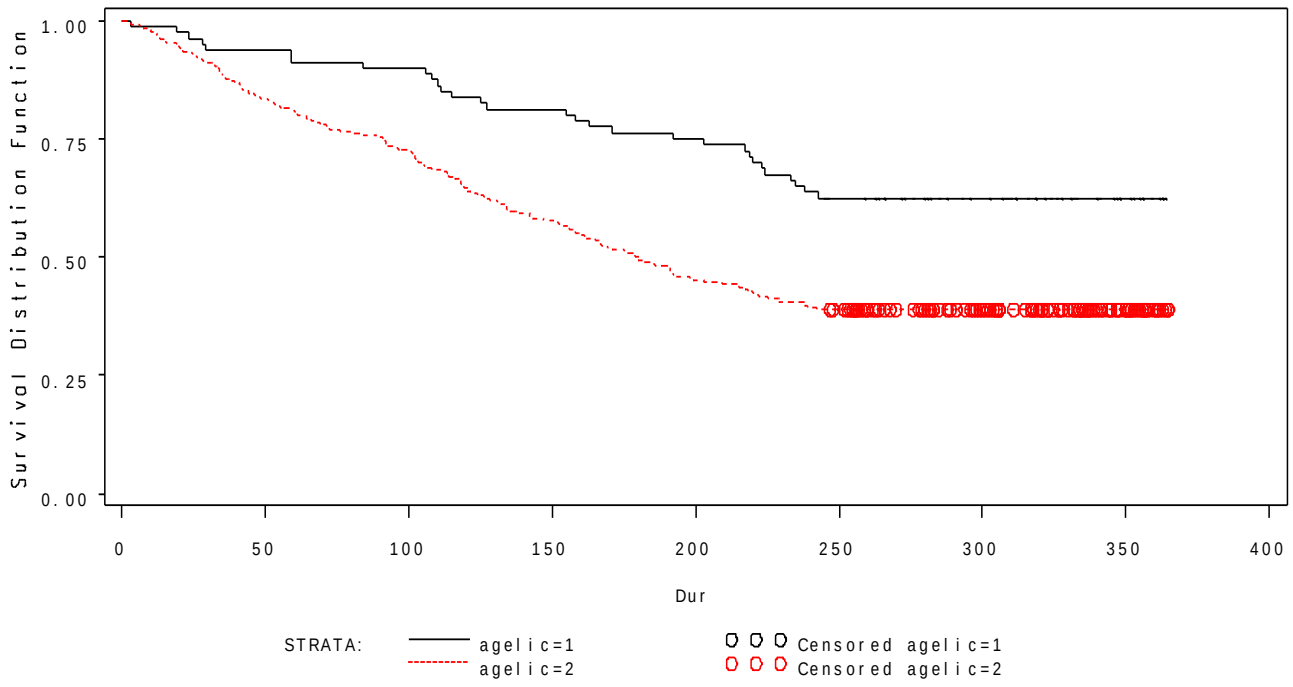

Figure 8. Survival distribution of the license duration of drivers.

moderate influence as seen in model 1A with p-value of 0.0807. Furthermore, in final model 1B, where the sex variable was artificially introduced yielded a p-value of 0.0408 and a hazard ratio ratio of 1.832 which means that the risk of female drivers had 1.832 times greater than the risk of male drivers. For survival distribution of

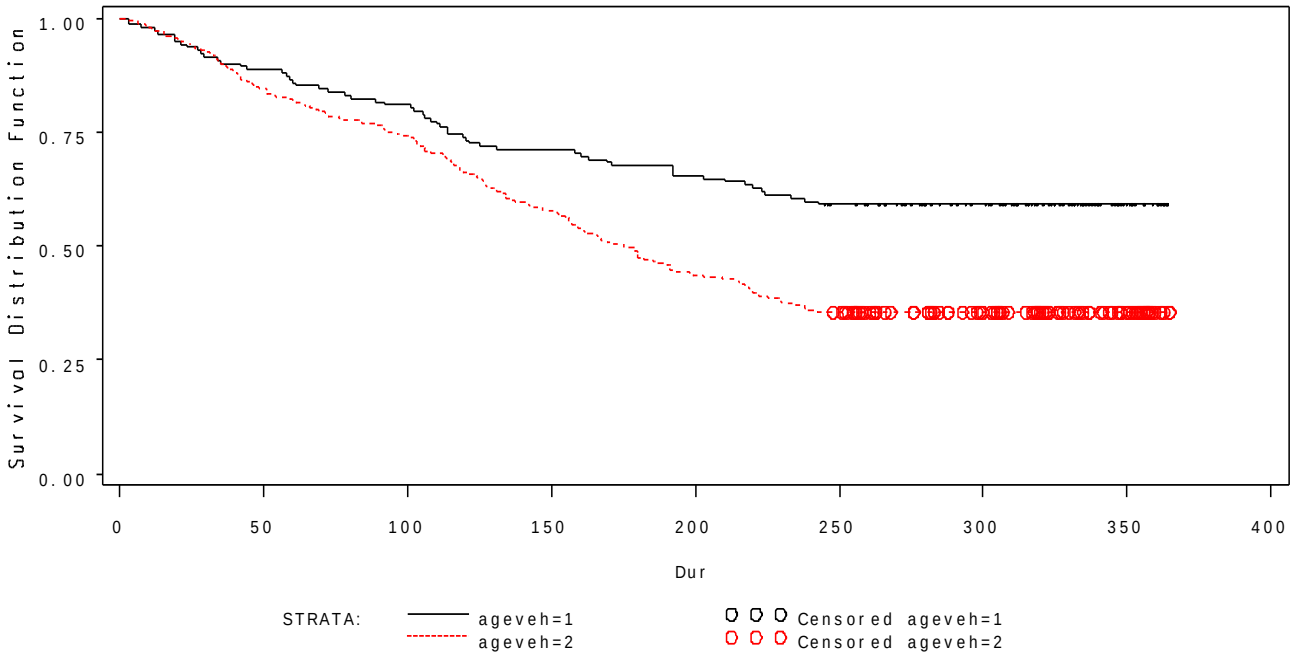

Figure 9. Survival distribution of ages of vehicles.

Table 2. Model 1A_ Kaplan-Meier proposed variables and other relevant variables.

The PHREG Procedure								
		Analysis of maximum likelihood estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**95% Confidence Limits Hazard Ratio (lower and upper limits)**	
sex	1	0.59775	0.34221	3.0510	0.0807	1.818	0.930	3.555
usebelt	1	-0.85539	0.25954	10.8622	0.0010	0.425	0.256	0.707
alcohol	1	1.10444	0.23274	22.5194	<.0001	3.018	1.912	4.762
ageveh	1	0.21207	0.23830	0.7920	0.3735	1.236	0.775	1.972
agelic	1	0.51500	0.27747	3.4448	0.0635	1.674	0.972	2.883
speedveh	1	1.18762	0.25391	21.8768	<.0001	3.279	1.994	5.394
tyrescon	1	0.48967	0.20705	5.5932	0.0180	1.632	1.087	2.448
wghtveh	1	-0.02177	0.18038	0.0146	0.9039	0.978	0.687	1.393
rutfam	1	-0.44185	0.25370	3.0332	0.0816	0.643	0.391	1.057
driver_agp2	1	0.61749	0.24996	6.1025	0.0135	1.854	1.136	3.027
driver_agp3	1	0.77469	0.42432	3.3332	0.0679	2.170	0.945	4.985
annukil2	1	-0.54953	0.25281	4.7249	0.0297	0.577	0.352	0.947
annukil3	1	-0.80285	0.32834	5.9788	0.0145	0.448	0.235	0.853

driver sex as shown in Figure 4, one can say that at any point in time the proportion of drivers estimated to be alive (not involved in an accident) is greater for males (represented by the upper curve) than that of the females (represented by the lower curve).

Generally, in the MTTD data, female drivers drove fewer kilometres, were less often under the influence of alcohol, used safety belt more often than male drivers, had fewer accidents and committed fewer offences. Previous studies demonstrated that differences in risks between sexes could be explained by differences in mobility. Therefore, the role of sex as a risk factor is less conclusive but may remain a practically useful explanatory variable.

Table 3. Linear hypotheses testing results.

Label	Wald Chi-Square	DF	Pr > ChiSq
driver_agp	6.4771	2	0.0392
annukil	6.5258	2	0.0383

Table 4. Mode 1B: Re-fitted significant variables in Model 1A.

The PHREG Procedure								
	Analysis of maximum likelihood estimates							
Variable	**DF**	**Parameter Estimate**	**Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**95% Confidence Limits Hazard Ratio (lower and upper limits)**	
sex	1	0.60565	0.29601	4.1862	0.0408	1.832	1.026	3.273
usebelt	1	-0.79550	0.22300	12.7254	0.0004	0.451	0.292	0.699
alcohol	1	0.91737	0.19118	23.0260	<.0001	2.503	1.721	3.640
speedveh	1	1.29852	0.22201	34.2100	<.0001	3.664	2.371	5.661
tyrescon	1	0.33968	0.18380	3.4153	0.0646	1.404	0.980	2.014
driverage	1	0.38741	0.15863	5.9641	0.0146	1.473	1.079	2.010
annukil	1	-0.43520	0.16507	6.9511	0.0084	0.647	0.468	0.894

Table 5. Model 1.1B_ Proportionality assumption test of final Model 1B.

Variable	Analysis of maximum likelihood estimates					
	DF	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
usebelt	1	-0.17771	0.98734	0.0324	0.8572	0.837
alcohol	1	0.58294	0.88965	0.4294	0.5123	1.791
speedveh	1	0.89538	1.00574	0.7926	0.3733	2.448
tyrescon	1	0.07544	0.85459	0.0078	0.9297	1.078
driver_agp	1	1.55025	0.76675	4.0878	0.0432	4.713
annukil	1	-1.29483	0.72598	3.1811	0.0745	0.274
usebeltt	1	-0.14984	0.21901	0.4680	0.4939	0.861
alcoholt	1	0.06074	0.19578	0.0963	0.7564	1.063
speedveht	1	0.09174	0.22128	0.1719	0.6784	1.096
tyrescont	1	0.07046	0.19135	0.1356	0.7127	1.073
driver_agpt	1	-0.26664	0.16767	2.5287	0.1118	0.766
annukilt	1	0.16893	0.16204	1.0868	0.2972	1.184

Driver's use of alcohol

Driving under the influence of alcohol significantly increased drivers accident risks. According to the final model 1B, alcohol use was a significant variable with p-value 0.0001 and the relative risks of drivers under the influence of alcohol, was 2.5031 times (1.721 – 3.640 with 95% confidence interval) greater at risk than that of non-alcohol users. From the survival distribution curve of use of alcohol as indicated in Figure 6, it can be seen that the survivorship function for non alcohol users is lying above the survivorship function for drivers who used alcohol which means that non alcohol users live longer or had a more favourable survival experience than the alcohol users.

Use of safety belt

Use of safety belt had a very strong explanatory power in model 1B with a p-value of 0.0004 and a hazard ratio of 0.451, indicating that if a driver changes from not use of belt to use of belt, while holding other covariates constant the hazard of accident decreases by (100 - 45.1%) = 54.9%. The use of safety belt, which was only supposed to influence the seriousness of injuries, proved to be also

Table 6. Linear hypotheses testing results.

Label	Wald Chi-Square	DF	Pr > ChiSq
test_proportionality	4.7929	6	0.5706

a strong risk factor in the models. From the survival graph of use of safety belt as seen in Figure 5, it can be seen that drivers who used safety belts had higher survival experience than those who did not. Aside its significant effect on accident time, it also increases the severity and consequences of the accident.

Speed of vehicle

Speed proved to be a statistically significant variable in predicting the hazard of accidents according to Kaplan-Meier estimate and the fitted Cox regression models. According to model 1B, the hazard of accidents for drivers who drove over 80 km/h is 3.664 times (with a 95% confidence interval: 2.371 – 5.661) that of those who drove less than 80 km/h. Also, from the survival distribution curve of estimated speed of vehicle at the time of accident as seen in Figure 7, it can be seen that drivers whose speeds exceeded 80 km/h had lower survival experience than those who did not.

Age of license

Age of license of drivers was used as a proxy to assess the level of experience of the driver. It was a moderate significant variable in model 1A with p=0.0635. However, the hazard ratio indicated that drivers with duration of license less than 5 years had 1.7 times the risk of those with license duration of at least 5 years. From the survival curve of driving experience (age of driving license in years) as indicated in Figure 8, it can be seen that drivers whose license age exceeded 5 years had higher survival experience than those with license less than 5 years.

Annual vehicle kilometreage

Drivers annual vehicle kilometreage had a strong explanatory power in the fitted models. In model 1A, it can be seen that drivers that had travelled between 5,000 km/a to 14,000 km/a had 42.3% lower than those that had travelled less than 5,000 km/a (the reference group). Also, those that had travelled for at least 15,000 km/a had (100 - 44.8%) = 55.2% lower than those that had travelled less than 5,000 km/a. In the final model 1B which gave the overall contribution of drivers' annual kilometreage (p=0.0084), the hazard ratio of 0.647 which means that for any one year increase in driver's exposure to traffic, it is associated with (100 - 64.7%) = 35.3% decrease in expected time to accident holding all other covariates constant. This results indicated that drivers accident risks decreases as annual kilometerage increases, but it is generally believed that higher exposure to traffic is associated with higher risk; besides, the findings of this research is in opposition to this belief. This perhaps might be due to accumulated experience on the part of these drivers.

Route familiarity

The route familiarity variable was not a significant variable as indicated in model 1A (p = 0.0816). The hazard ratio is 0.643 indicating that, the accident risks was about 35.7% lower for drivers who were familiar with the site of the accident compared to other drivers.

Vehicle weight

The statistical significance of vehicle weight in model 1A was very weak (p =0.9039) and therefore its role as a risk factor is less conclusive. However, users of light vehicles had a higher relative risk.

Vehicle age

Vehicle age had no significant effect on accident risk in the model 1A, though it proved to be a significant variable in the Kaplan-Meier estimate. In this model, the hazard ratio is 1.236, indicating that vehicles older than 10 years had 1.2 times the risk of vehicles that are less or equal to 10 years. In otherwords, for each one year increase in the the age of the vehicle, the hazard of accident goes up by estimated 100 (1.236 – 1) = 23.6%. Also, from the survival curve of age of vehicle as shown in Figure 9, it can be seen that drivers whose vehicles aged over 10 years had lower survival experience than those with vehicles less than 10 years. However, it was observed in the MTTD data that users of old vehicles included many alcohol users, young drivers, and non-users of belt. More drivers of newer vehicles were involved in speeding over 80 km/h prior to the accident.

Tyres condition / Tread depth

Tyres condition/tread depth proved to be statistically significant (p = 0.018) risk factor in determining accident time in model 1A but had a moderate influence in the final

Table 7. Data set A: Diagnostic for Influential observations of final model 1B.

obs	Dur	Status	Sex	Usebelt	Alcohol	speedveh	tyrescon	driver_agp	annukil	dsex	dusebelt	dalcohol	dspeedveh	dtyrescon	Ddriver_agp	dannukil
1	35	1	1	0	1	.	0	2	2	.	.	.	.	.	.	.
2	36	1	1	0	1	2	0	2	3	0.012141	0.001596	0.003132	-0.001888	0.001891	-0.000773	0.020836
3	51	1	1	0	1	2	0	2	2	-0.003374	-0.002846	0.002970	0.001920	-0.005749	0.000558	-0.002157
4	83	1	1	0	1	2	0	2	3	0.011250	0.001679	0.002897	0.000025	0.001193	0.000154	0.016048
5	102	1	1	0	1	1	0	2	2	-0.002860	-0.011235	0.007426	-0.037988	-0.009936	-0.002045	0.000163
6	105	1	1	.	1	2	0	2	3	.	.	.	.	.	.	.
7	112	1	1	0	1	2	0	2	2	-0.000401	-0.001800	0.001805	0.002220	-0.003833	0.001072	-0.002574
8	118	1	1	0	1	2	1	2	2	0.002781	-0.000429	0.000673	0.001172	-0.001904	0.000436	-0.001596
9	6	1	1	0	1	2	0	2	2	-0.005225	-0.005328	0.003429	0.001740	-0.008330	-0.000824	-0.003104
10	12	1	1	0	1	2	1	2	2	-0.010517	-0.002869	0.004401	0.005014	0.023395	0.001371	0.006249
11	15	1	1	.	0	2	0	2	3	.	.	.	.	.	.	.
12	19	1	1	1	1	2	1	2	1	-0.022978	0.036582	0.008396	0.017079	0.016261	0.006785	-0.012530
13	21	1	1	0	1	2	0	2	2	-0.005470	-0.004847	0.003208	0.001747	-0.007522	-0.000584	-0.003031
14	23	1	1	0	1	2	1	2	2	-0.010135	-0.002646	0.004186	0.004614	0.021784	0.001229	0.005722
15	26	1	1	0	1	1	1	2	1	-0.028411	-0.017256	0.009033	-0.036386	0.011288	-0.001437	-0.014849
16	31	1	1	.	1	1	0	2	1	.	.	.	.	.	.	.
17	34	1	1	0	1	2	0	2	2	-0.004707	-0.003824	0.003259	0.001760	-0.006226	-0.000107	-0.002565
18	40	1	1	.	1	1	0	2	2	.	.	.	.	.	.	.
19	42	1	1	0	0	2	0	2	2	-0.013473	-0.007866	-0.029537	0.006110	-0.008439	0.001386	-0.002555
20	47	1	1	0	1	2	0	2	2	-0.003473	-0.002912	0.003039	0.002061	-0.005899	0.000574	-0.002224
21	60	1	1	1	1	2	1	2	1	-0.015668	0.026717	0.006300	0.012945	0.011399	0.005719	-0.009193
22	72	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
23	77	1	1	1	1	2	1	2	.	.	.	.	.	.	.	.
24	69	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
25	91	1	1	.	1	2	0	2	2	.	.	.	.	.	.	.
26	94	1	1	0	0	2	0	2	2	-0.010835	-0.006641	-0.025588	0.006906	-0.007427	0.001802	-0.003015
27	103	1	1	0	1	2	0	2	3	0.009523	0.001238	0.002647	-0.000100	0.000871	0.000503	0.014004
28	106	1	1	0	1	2	0	2	1	-0.000606	-0.000313	0.000549	0.001205	-0.001088	0.000891	-0.001811
29	114	1	1	.	1	1	1	2	2	.	.	.	.	.	.	.
30	10	1	1	0	1	2	0	3	2	-0.000924	-0.000017	0.002038	0.005150	-0.006219	0.022966	-0.003713
31	28	1	2	1	1	2	1	2	1	0.045208	0.033748	0.013548	0.012590	0.009208	0.008729	0.004491
32	59	1	2	0	0	2	1	1	1	0.026815	-0.006971	-0.012306	0.006716	0.003847	-0.009713	0.000049
33	108	1	2	0	1	2	0	2	1	-0.048687	0.003980	-0.010258	-0.003753	0.018194	-0.003365	0.007634
34	119	1	1	.	1	2	1	2	1	.	.	.	.	.	.	.
35	120	1	1	0	0	2	0	2	2	-0.007739	-0.006798	-0.021860	0.006877	-0.007676	0.001811	-0.003833
36	92	1	1	1	1	1	1	2	2	-0.002184	0.035521	0.012881	-0.030291	0.022223	0.004101	0.015555

Table 7. Contd.

37	124	1	1	0	1	2	0	1	2	-0.002394	-0.005544	0.003366	0.001340	-0.005900	-0.011137	-0.002885
38	125	1	1	0	1	2	0	2	2	0.000930	-0.001771	0.001582	0.002152	-0.002671	0.000458	-0.002164
39	131	1	1	0	0	2	0	1	3	0.006272	-0.007238	-0.025665	0.001328	-0.002797	-0.020053	0.019385
40	134	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
41	135	1	1	0	1	2	0	2	.	.	.	.	.	.	.	.
42	142	1	1	1	1	1	1	2	1	-0.015493	0.022896	0.011892	-0.020449	0.010138	0.003974	-0.008564

Table 8. Interesting and important variables in the survival models of the MTTD data

Variable	Remark
Driver Age	The relative risk was high for both the middle aged and the old but lowest for the young
Driver Sex	Female drivers had a higher relative risk than male drivers.
Use of alcohol	Significantly increased drivers' relative risk
Not using the safety belt	Significantly increased drivers' relative risk
Speed	Higher relative risk for drivers with speed over 80km/h
Familiarity of route	Drivers' familiar with route had a lower relative risk
Annual vehicle kilomreage	Relative risk decreased with increase in kilometerage
Vehicle Age	Users of new vehicles had somewhat lower risk, (though not statistically significant variable).
Vehicle weight	Users of light vehicles had a higher relative risk, (though not statistically significant variable).
Tyres tread depth	Small tread depth increases relative risk (though not statistically significant variable)
Age of license	Drivers with lessthan 5 years of license duration had higher relative risk.

model with p=0.0646. Therefore, the role of tyres condition as a risk factor is less conclusive. However, users of worn out tyres are more at risk than users of unworn out tyres. According to the final model 1B, drivers with very worn out tyres vehicles had 1.404 times that of drivers who used less worn out tyres (>4 mm). This also means that for every 1 mm decrease in the tyre's tread depth, the accident probability increased by about 40%.

Assessing the adequacy of the final model 1B

The proportionality assumption test of the model is indicated in model 1.1B. It can be seen that the tests of all the time-dependent variables were not significant either individually (Table 5) or collectively (Table 6) with p-value for each variable greater than 0.05. Therefore, we do not have enough evidence to reject proportionality assumption for this model. Since the assumption of proportionality is satisfied, it suggests that the Cox regression model 1B provides a reasonable fit to the MTTD data.

The DfBeta statistics was also employed to determine whether there are any influential observations in the data in the fitting of the model, which is displayed in Table 7. The table displays the Dfbeta statistics dataset for first 42 individuals

only; the signs of the DfBeta statistics are the reverse of what one might expect – a negative sign which means that the coefficient increases when the observation is removed. For example, the estimated coefficients for the covariates of sex and alcohol in the final model are respectively 0.60565 and 0.91737. However, in Table 7, the value 0.012141 for *dsex* indicates that if observation 2 is removed, the sex coefficient will decrease to approximately 0.60565-0.012141 = 0.593509, a decrease of 2%. Also, the value 0.0013132 for *dalcohol* indicates that if observation 2 is removed, the alcohol coefficient will decrease to approximately 0.91737-0.003132 = 0.914238, a decrease of 0.3%. Furthermore, it can be seen that the value -0.015493 for *dsex* indicates that if observation 42 is removed, the sex coefficient will increase to approximately 0.60565+0.015493 = 0.621143, an increase of 1.5%. Overall, it can be seen that none of the observations did exert an undue influence on the estimated coefficients and hence the fit of the model.

In summary, the evaluation of the final model based on the proportionality assumption and the DfBeta indicated that the models structure was acceptable. The practical conclusion is that removal of an observation will result in no or minor changes in the overall coefficients of all the covariates considered and hence will not distort the models. However, missing data was a general problem. The amount of missing data varies from one variable but the models included all the observations that did not have missing data concerning the variables. Specific causes of missing data were drivers who had died in the accidents and could not be captured in the registry. Some killed drivers were omitted from the MTTD data due to large missing information about them.

Conclusion

It is clear from the analysis that the application of survival models to the analysis of accident data appeared to be a promising approach. The models applied well to the examination of accident risk factors. Table 8 presents the most important risk factors according to the models for the MTTD accident data.

RECCOMMENDATIONS

1) This study focused on only one year driving, but the driver might have been driving for so many years before the accident occurred. Therfore, any further work on this should be focused on the date the drivers got their driving license to their first accident involvement.
2) There are several covariates that may play an essential role to the development of the models but unfortunately were not available. Example, the date the vehicle was first taken into use, how long the driver had been on the trip when accident occurred, criminal records of drivers, drivers history of accident involvement, and income levels of drivers,.
3) Statkeholders responsible for ensuring safety on our roads should implement the findings of the study since it will enable them put up better measures to reduce the occurrence of accidents in the northern region in particular and the country as a whole.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES

Allison PD (1995). Survival Analysis using SAS: A practical guide, Cary, NC: SAS Institute Inc.

Blower D, Kenneth L, Campbell, Green P (1993). Accident rates for heavy truck-tractors in Michigan. Accid. Anal. Prev. 25(3):307-321.

Cain KC, Lange NT (1984). Approximate case influence for the proportional hazards regression model with censored data. Biometrika 40:493-499.

Chieng-Ming T, Ming-Shan Y, Li-Yung T, Hsin-Hsien L, Min-Chi L (2016). A comprehensive analysis of factors leading to speeding offenses among large-truck drivers. Transp. Res. Part F. 38:171-181.

Calliendo C, De-Guglielmo ML, Guida M (2013). A crash-prediction model for road tunnels. Accid. Anal. Prev. 55:107-115.

Collett D (2003). Modelling Survival Data in Medical Research, Chapman and Hall, London.

Cox DR, Snell EJ (1968). A general definition of residuals with discussion. Royal Statistical Society J. Series B. 30:248-275.

Cox DR (1972). Regression models and life-tables. Royal Statistical Society J. Series B. 34:187-220.

Cox DR (1975). Partial likelihood. Biometrika 62:269-276.

Crowley J, Hu M (1977). Covariance analysis of heart transplant survival data. Am. J. Stat. Ass. 78:27-36.

Dagan Y, Doljansky JT, Green A, Weiner A (2006). Body mass index (BMI) as a first line -screening criterion for detection of excessive daytime sleepiness among professional drivers. Traffic Injury Prev. 7(1):44-48.

Elvik R (1996). A Metaanalysis of studies concerning the safety effects of daytime running lights on cars. Accid. Anal. Prev. 28(6):685-694.

Elvik R, Christensen P, Amundsen A (2004). Speed and road accidents: an evaluation of the power model. Institute of Transport Economics (TØI) Report. 740:134.

Hakkanen J, Summala H (2001). Fatal traffic accidents among trailer truck drivers and accident causes as viewed by other truck drivers. Accid. Anal. Prev. 33:187-196.

Häkkänen J, Summala H (2000). Sleepiness at work among commercial truck drivers. Sleep 23(1):49-57.

Jovanis PP, Kaneko T, Lin TD (1991). Exploratory analysis of motor carrier accident risk and daily driving patterns. Transp. Res. Board. Working paper No. 73.

Kalbfleisch JD, Prentice RL (2002). The Statistical Analysis of Failure Data, 2nd ed. Wiley, New York.

Kaplan E, Meier P (1958). Nonparametric estimation from incomplete observations. Am. J. Stat. Ass. 53:457-481.

Klein JP, Moeschberger ML (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York.

Klembaum DG (1996). Survival Analysis: A Self learning text. Springer, New York.

Lawless JF (1982). Statistical Models and Methods for Lifetime Data Analysis. Wiley, New York.

Miaou SP (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26(4):471-482.

Oppe S (1989). Macroscopic models for traffic and traffic safety. Accid. Anal. Prev. 21(3):225-232.

Rodríguez DA, Rocha M, Khattak AJ, Belzer MH (2003). Effects of truck driver wages and working conditions on highway safety: Case study. Trans. Res. Record 1833:95-102.
Smeed RJ (1949). Some statistical aspects of road safety research. Journal of Royal Statistical Society, Series A. 112(1):1-34.
Schoenfeld D (1982). Partial residuals for the proportional hazards regression model. Biometrika 69:239-241.
Saccomanno F, Buyco C (1988). Generalized Loglinear Models of Truck Accident Rates. Trans. Res. Record 1172:23-31.
Sullman JM, Meadows ML, Pajo KB (2002). Aberrant driving behaviours amongst New Zealand truck drivers. Trans. Res. Part F Traffic Psychol. Behav. 5:217-232.
Taylor AH, Dorn L (2006). Stress, fatigue, health, and risk of road traffic accidents among professional drivers: The contribution of physical inactivity. Annual Rev. Public Health 27:371-391.
Thernaeau TM, Grambsch PM, Fleming TR (1990). Martingale-based residuals for survival models. Biometrika 77:147-160.