HÜLYA OLMUŞ, Ph.D.
E mail: hulya@gazi.edu.tr
SEMRA ERBAŞ, Ph.D.
E mail: serbas@gazi.edu.tr
Gazi University, Faculty of Sciences
Department of Statistics
Teknikokullar, Ankara, Turkey

Human - Transport Interaction
Preliminary Communication
Accepted: Dec. 28, 2011
Approved: Nov. 13, 2012

ANALYSIS OF TRAFFIC ACCIDENTS CAUSED BY DRIVERS BY USING LOG-LINEAR MODELS

ABSTRACT

Log-linear modelling is advanced as a procedure to identify factors that underlie the relative frequency of occurrence of various characteristics. The purpose of this study is to present a modelling effort using log-linear models to estimate the relationships between driver's fault and carelessness and the traffic variables such as gender, accident severity, and accident time. The study was conducted in four different districts in Ankara, the capital of Turkey. There were 1,325 people selected for the study; and they were asked whether they had been in an accident. Four hundred and forty-eight of them answered that they had been involved in an accident. As drivers, 276 out of 448 people, namely 61.6%, had traffic accidents. The data on the variables, namely gender, driver's fault and carelessness, accident severity and accident time, were collected through a questionnaire survey. Detailed information has been created based on this information. The analysis showed that the best-fit model regarding these variables was the log-linear model. Furthermore, the odds ratio between these variables, the associations of the factors with the accident severity and the contributions of various factors, and the multiple interactions between these variables were assessed. The obtained results provide valuable information in regard to preventing undesired consequences of traffic accidents.

KEYWORDS

Log-linear model, traffic accidents, odds ratio, likelihood-ratio test statistics

1. INTRODUCTION

An increased number of vehicles on the roads has been in parallel to the growth of the automotive sector in the last years. Also, traffic accidents have increased in parallel to increased vehicles. Turkey is lagging behind many European and other countries in terms of vehicle numbers, but also in the upper levels regarding traffic accidents.

Road traffic accidents result in the deaths of more than 500 thousand people and injury of many more throughout the world annually. According to the Ministry of Health data in Turkey, death caused by traffic accidents ranks third among all the known causes of death. More than 4 million traffic accidents have occurred in the most recent ten-year period in Turkey, and an average of 5 thousand people lost their lives annually because of these accidents. Moreover, an average of 750 thousand accidents, with loss of lives, injury, and/or economic loss, occur annually in the country. As for the statistics regarding population, one in every seven people has been killed or injured in a traffic accident, or had a relative who suffered in an accident. It is clear that more emphasis should be placed on the concept of traffic accidents, since the problems caused by them are ever increasing in seriousness.

Many statistical methods are used in research related to traffic accidents, one being the log-linear model. Several papers have used log-linear models for studying traffic accidents, but this method has been used in a limited way, even though it gives valuable results and provides information on multiple interactions between various factors which are useful for understanding the overall problem. Kim et al. (1995a) used accident type, seat-belt use, and injury severity variables to find the relationship among these three factors. They found that there is a relationship between crash types and injury levels: the most serious injury producing collisions involve head-on and rollover collisions. Kim et al. (1995b) estimated a log-linear model to investigate the role of driver characteristics and behaviours in the causal sequence leading to more severe injuries. They found that the younger driver is more likely to be classified as fatal than the older driver involved in accidents. Richardson et al. (1996) used the log-linear model to study the patterns of motor vehicle accident involvement as regards driver age and gender in Hawaii. They found that young drivers have a much greater frequency of roll-overs and of being the rear-ender or head-oner, whereas older drivers have a much higher frequency of being rear-ended or side-swiped. Abdel-Aty et al. (1998) used the log-linear model to estimate the relationships between driver age and several significant factors in the multivariate context. They found that there are significant relationships between the driver age and average daily traffic, injury severity, manner of collision, speed, alcohol involvement, and roadway character. That is, their findings reveal that injury severity is related to age, and the old and very old drivers are more likely to be killed in traffic accidents probably due to the decline in their physical condition. In terms of driver effects, Lourens et al. (1999) have concluded that there is no difference between men and women in terms of their crash involvement – after controlling for annual miles driven. They found younger drivers have the highest crash involvement rate per mile-driven among all age groups and a recent history of drinking violations have a positive effect on fatal crash rates. Rather interestingly, they also concluded that education level is irrelevant to crash involvement. Jang (2006) used the model to estimate the relationships between driving behaviour and driver's characteristics. He found there is mainly a relationship between gender and traffic accidents, but no relationship between education level and traffic accidents.

In this study, drivers who had been involved in traffic accidents in Ankara, the capital of Turkey, were studied. The variables used in the analysis presented in this paper are related to the gender of the driver, the time of the accident, accident severity, and driver's fault and carelessness on the roads. It was not certain whether these factors played a role in the outcomes of the traffic accident. Such information concerning these factors is important because it shows that preventive decisions can be carried out in regards to driving traffic. Much spending and education undertaken over many years have not been sufficient for Ankara. Therefore, a log-linear analysis was used for determining the relationship between these variables, and an attempt was made to develop a model that is a best-fit for them. Also, odds ratios and the contributions of various factors, and multiple interactions between variables were obtained.

2. DATA PRESENTATION AND RESEARCH HYPOTHESIS

The data reported in this study were collected as part of an extensive questionnaire survey in Ankara, the capital of Turkey, during the period from December 2007 to December 2008. A total of 1,325 drivers were chosen for this analysis; 276 of them had a traffic accident. The sample units used in this study were formed using the data of the 2007 Population Census as the basis. The data of this Census that had been categorized by the criteria of age and gender was used in the study. The stratified sampling method was used; and the stratification units were formed from the four large districts of Ankara. The determined sample was distributed according to age and gender; and the people who would be presented with the questionnaire were determined.

There were 276 individuals who had traffic accidents as drivers, taken for the analysis. Therefore, four variables were summarized from the data. The description and levels of these variables are given in *Table 1*.

The driver's fault and carelessness variables have two categories. Determination of driver's fault and carelessness depends on the drivers' statements. An accident which has not happened due to the driver's fault and carelessness has been accepted as road fault, other driver's fault, pedestrian fault, etc.

The accident time variable was classified into 3 groups: The first category of accident time non-holiday time variable includes daytime accidents occurring on weekdays and weekends while the second category, termed night, includes night-time accidents on weekdays and weekends. The daytime represents the time period between 7.00 a.m. to 6.00 p.m. The third category, termed holiday, includes accidents occurring during summer vacations, and religious and national holidays.

Table 1 - Description of variables

Number	Variables	Coding/values	Abbreviations
1	Sex	1=Male 2=Female	SEX (S)
2	Driver's fault and carelessness	1=Yes 2=No	DRIVER (D)
3	Accident time	1=Daytime 2=Night-time 3=Holiday time	TIME (T)
4	Accident severity	1=Injury/death 2=Without injury/death (Economic losses)	RESULT (R)

Since this distinction between categories of accident is not as important as the night-time or daytime perspective, these categories are mutually exclusive.

The accident severity variable was classified into two groups: with injury/ death and economic losses. The injury/death category includes strain, twist, and other injury cases while those without injury/death category include economic losses. As before, these also depend on the declaration of the drivers who had the accident.

In this study, the null hypothesis which claims that there is no relationship between the accident severity and other variables such as sex, the driver's fault and carelessness and accident time, respectively, will be tested against the research hypothesis. In testing the null hypothesis p-value is used. If the p-value is less than the significance level, we reject the null hypothesis and the data are said to be "statistically significant" at level α [8].

3. STATISTICAL METHODS

The use of log-linear modelling has been recommended as a statistical analysis method when the dependent and independent variables are categorical in nature [4]. An important advantage that log-linear modelling has over other techniques such as analysis of variance (ANOVA), chi-square or variance accounted for procedures, is its statistical power. A unique feature of log-linear analysis is its ability to capture the interrelationships among subjects' responses and the factorial structure of the study design categories [3]. A log-linear model describes the association and interaction patterns among a set of categorical variables.

In practice, we try to fit a model so as to avoid using a saturated model. The saturated model in log-linear analysis is a kind of model that incorporates all the possible effects, such as one-way effect, two-way interactions effect, three-way interactions, etc. A saturated model imposes no constraints on the data and always reproduces the observed cell frequencies. The parsimonious models in log-linear analysis are incomplete models that somehow achieve a satisfactory level of goodness of fit.

Log-linear analysis deals with the association of categorical or grouped data, looking at all levels of possible main and interaction effects and comparing this saturated model with reduced models, with the primary purpose being to find the most parsimonious model which can account for cell frequencies in a table. That is, log-linear analysis is an independent procedure for accounting for the distribution of cases in a joint distribution or cross tabulation of categorical variables.

The log-linear model is generally called a hierarchical model: this means that whenever the model contains higher-order effects, it also incorporates lower-order effects composed of the variables. For instance, when the model contains, λ_{ij}^{SD}, which is an interaction of S at the i^{th} level and D at the j^{th} level, it also must contain λ_i^S, effect due to the i^{th} level of S, λ_j^D, effect due to the j^{th} level of D. The reason for including lower-order terms is that the statistical significance and practicable interpretation of a higher-order term depend on how the variables are coded. This is undesirable, but with hierarchical models, the same results are obtained, irrespective of how the variables are coded [2].

As the number of dimensions of a contingency table increases, the number of possible models also increases. Hence, some procedures are clearly needed to indicate which model may prove reasonable for the data set and which are likely to be inadequate. One such procedure is to examine the likelihood-ratio chi-square values of all effects in the saturated log-linear model. The other approach is to examine the standardized parameter values in the saturated log-linear model. These values may indicate which unsaturated models may be excluded, and consequently which unsaturated models may be worth considering [11].

Log-linear analysis uses a likelihood-ratio chi-square (G^2), and therefore can be used to analyze greater than two-way tables. To determine whether there is a significant difference between the observed and expected frequencies, the likelihood–ratio chi-square (G^2) is computed. The definition of the likelihood-ratio chi-square is:

$$G^2 = 2\sum\left[O_i \ln(O_i/E_i)\right]$$

where O_i is the observed frequency and E_i is the expected frequency.

The log-linear model for the contingency table is expressed as follows:

Log(expected cell frequency) is a sum grand mean, main effects parameters and second and higher order interactions.

For example, the model with only four main effects, that is, the independent model, is,

$$\log m_{ijkl} = \mu + \lambda_i^S + \lambda_j^D + \lambda_k^T + \lambda_l^R,\ i = 1,2;\ j = 1,2;\\ k = 1,2,3;\ l = 1,2\cdots$$

where μ is overall effect; λ_i^S is effect due to the i^{th} level of S; λ_j^D is effect due to the j^{th} level of D; λ_k^T is effect due to the k^{th} level of T and λ_l^R is effect due to the l^{th} level of R. We impose the sum-to-zero identifiability conditions

$$\sum_i \lambda_i^S = \sum_j \lambda_j^D = \sum_k \lambda_k^T = \sum_l \lambda_l^R = 0$$

For example, the model with four main effects and six terms of two-way interactions, that is, the second order full model, is:

$$\log m_{ijkl} = \mu + \lambda_i^S + \lambda_j^D + \lambda_k^T + \lambda_l^R + \lambda_{ij}^{SD} + \lambda_{ik}^{ST} + \lambda_{il}^{SR} + \lambda_{jk}^{DT} +\\ + \lambda_{jl}^{DR} + \lambda_{kl}^{TR},\ i = 1,2;\ j = 1,2;\ k = 1,2,3;\ l = 1,2$$

where $\log m_{ijkl}$ is logarithm expected frequency of cell in which $S=i$, $D=j$, $T=k$, $R=l$; μ is overall effect; λ_i^S is effect due to the i^{th} level of *S*; λ_j^D is effect due to the j^{th} level of *D*; λ_k^T is effect due to the k^{th} level of *T*; λ_l^R is effect due to the l^{th} level of *R*; λ_{ij}^{SD} is interaction of *S* at the i^{th} level and *D* at the j^{th} level; λ_{ik}^{ST} is interaction of *S* at the i^{th} level and *T* at the k^{th} level; λ_{il}^{SR} is interaction of S at the i^{th} level and *R* at the l^{th} level; λ_{jk}^{DT} is interaction of *D* at the j^{th} level and *T* at the k^{th} level; λ_{jl}^{DR} is interaction of *D* at the j^{th} level and *R* at the l^{th} level; λ_{kl}^{TR} is interaction of *T* at the k^{th} level and *R* at the l^{th} level. We impose the sum-to-zero identifiability conditions

$$\sum_i \lambda_i^S = \sum_j \lambda_j^D = \sum_k \lambda_k^T = \sum_l \lambda_l^R = 0$$

$$\sum_i \lambda_{ij}^{SD} = \sum_j \lambda_{ij}^{SD} = \sum_i \lambda_{ik}^{ST} = \sum_k \lambda_{ik}^{ST} = 0$$

$$\sum_i \lambda_{il}^{SR} = \sum_l \lambda_{il}^{SR} = \sum_j \lambda_{jk}^{DT} = \sum_k \lambda_{jk}^{DT} = 0$$

$$\sum_j \lambda_{jl}^{DR} = \sum_l \lambda_{jl}^{DR} = \sum_k \lambda_{kl}^{TR} = \sum_l \lambda_{kl}^{TR} = 0$$

S, *D*, *T*, and *R* here are the abbreviations for the variables presented in *Table 1*. Therefore, in the section below an attempt is made to determine the best-fit model, as regards to the data set used in the study.

4. RESULTS

4.1 Basic Statistical Analysis

Two-way contingency tables are formed to calculate the conditional probabilities. In this paper, one variable is the variable categories of interest, termed as the row variables, and the other is the accident severity, termed as the column variable. Then it is informative to construct a separate probability distribution for the row variable at each level of the column variable. Such a distribution consists of conditional probabilities for the row variable, given the level of the column variable [5]. The study presents only the statistically significant results of the Chi-square test (χ^2). In *Table 2*, frequency distributions for sex are given.

Table 2 - Frequency distribution of accident severity and sex

Sex (S)	Accident severity (R)		Total
	Injury/death	Economic losses	
Male	56 (30.6%)	127 (69.4%)	183 (100.0%)
Female	45 (48.4%)	48 (51.6%)	93 (100.0%)
Total	101 (36.6%)	175 (63.4%)	276 (100.0%)

χ^2 = 8.407, *p-value* = 0.004

In terms of the accident severity variable, approximately 36.6% and 63.4% of the drivers are classified as injury/death and economic losses, respectively. The female drivers have a higher proportion (48.4%) of injury/death while the male drivers' share is slightly lower (30.6%). However, the male drivers have a higher proportion (69.4%) of economic losses. The null hypothesis of independence between the sex and accident severity is rejected at p-value<0.01. It is evident that the accident severity is highly related to sex. Frequency distributions for driver's fault and carelessness are contained in *Table 3*.

Table 3 - Frequency distribution of accident severity and driver's fault and carelessness

Driver's fault and carelessness (D)	Accident severity (R)		Total
	Injury/death	Economic losses	
Yes	55 (30.8%)	124 (69.2%)	179 (100.0%)
No	46 (47.4%)	51 (52.6%)	97 (100.0%)
Total	101 (36.6%)	175 (63.4%)	276 (100.0%)

χ^2 = 7.558; *p-value* = 0.006

Frequency distributions for driver's fault and carelessness are contained in *Table 3*. It is shown that the drivers who are at fault and careless have a higher proportion (69.2%) of economic losses while the drivers who are not at fault and careless are slightly lower (52.6%). The null hypothesis of independence between the driver's fault and carelessness and accident severity is rejected at p-value<0.01. It is evident that the accident severity is highly related to driver's fault and carelessness.

Frequency distributions for accident time are contained in *Table 4*.

Table 4 - Frequency distribution of accident severity and accident time

Accident time (T)	Accident severity (R)		Total
	Injury/death	Economic losses	
Daytime	26 (28.9%)	64 (71.1%)	90 (100.0%)
Night time	45 (47.4%)	50 (52.6%)	95 (100.0%)
Holiday time	30 (33.0%)	61 (67.0%)	91 (100.0%)
Total	101 (36.6%)	175 (63.4%)	276 (100.0%)

χ^2 = 7.572; *p-value* = 0.023

It is shown that the night time tends to have a higher injury/death proportion (47.4%), while the daytime share is slightly lower (28.9%). The null hypothesis of independence between the accident time and accident severity is rejected at p-value<0.05. It is evident that the accident severity is highly related to accident time.

When Tables 2, 3 and 4 are evaluated, sex, driver's fault and carelessness and accident time emerge as significant factors for accident severity.

A measure of association between the row categories and the column categories is *relative risk*; another measure is *odds ratio*. The odds ratio is a way of comparing whether the probability of a certain event is the

Table 5 - Relative risk , odds ratio, and 95% confidence intervals for odds ratio for the variable of accident severity

Variable	Relative risk	Odds ratio	95% Confidence Interval for Odds ratio
Sex	0.764	0.470	0.281-0.786
Driver's fault and carelessness	0.769	0.492	0.295-0.819
Daytime- Night time	0.652	0.451	0.246-0.829
Daytime-Holiday	0.907	0.826	0.439-1.553
Night time-Holiday	1.332	1.830	1.010-3.316

same for two groups. The odds ratio is a measure of effect size, describing the strength of association or non-independence between two groups. An odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than one implies that the event is more likely in the first group. An odds ratio less than one implies that the event is less likely in the first group.

A more direct measure comparing the probabilities in two groups is the relative risk, which is also known as the risk ratio. The relative risk is the ratio of the proportions of cases having a positive outcome in the two groups. Like the odds ratio, a relative risk equal to one implies that the event is equally probable in both groups. A relative risk greater than 1 implies that the event is more likely in the first group. A relative risk less than 1 implies that the event is less likely in the first group.

For the *relative risk and odds ratio*, according to *Table 5*, some of the interpretations below can be made:

- *The relative risk* of injury/death is 0.764; in other words, there is a 0.764 greater probability of injury/death for males than for females.
- *The relative risk* of injury/death is 0.769 for the variable of driver's fault and carelessness. Thus, there is a 0.769 greater probability of injury/death for driver's fault and carelessness than without driver's fault and carelessness.
- *The odds ratio* is 1.830 between the categories of night time and holiday. In other words, there is nearly a twofold greater odds of injury/death for night time than for holiday time.

4.2 Application of Log-linear analysis

The relation between variables has been analyzed by log-linear analysis. For the analysis of data, the SPSS 15.0 package program was used. In this study, we first searched for the simplest relationship among variables.

Table 6 gives an initial idea of what order(s) of effects are or are not appropriate for the most parsimonious model. In *Table 6* the column labelled 'p-value' gives the observed significance levels for the tests where K-way and higher order effects are zero. The K factor relates to the number of interactions in the classification *Table 6*. A small observed significance level indicates the hypothesis that terms of particular orders of zero should be rejected. It is clear in *Table 6* that interaction terms up to the second order are sufficient to explain the variations in observed cell frequencies.

Table 6 - Tests whose K-way and higher order effects are zero

K	df	G^2 (Likelihood ratio)	p-value
4	2	1.355	0.508
3	9	11.870	0.221
2	18	61.503	0.000
1	23	136.366	0.000

In *Table 6*, the first line, K=4, gives the G^2 for the model without the four-factor interaction S*D*R*T. That is, the line tests the hypothesis that SDRT=0. The line with K=3 indicates the model without the fourth and third order effects because of the hierarchy principle. That is, the line tests the hypothesis that SDR=SDT=SRT=DRT=0. From these results, there is no sufficient reason not to accept these hypotheses for K=3 and K=4 (p>0.05). Similarly, K=2 indicates the model without the fourth, third and second effects. That is, the line tests the hypothesis that SD=ST=SR=...=0. The line, K=1, corresponds to a model that has no effects. The first and second rows' effects were significant (p<0.05). Finally, a model with the first and second order effects would seem adequate to represent our data. The degrees of freedom (df) for K=3 is the sum of degrees of freedom corresponding to all three-way and four-way interactions (such as, the number of parameters for the S*R*T interaction term equals (2-1)*(2-1)*(3-1)=2).

Testing for which effects are significant used Partial Chi-Square statistics. The effects of the first and the second row of the Partial chi-square analysis outcomes are given in *Table 7*. The tests of partial association are partial likelihood ratio tests and are based on the difference in likelihood of the ratio chi-square for the model with and without a given term.

In *Table 7* the model run showed that *time*sex*, *time*driver*, *result*sex*, *result*driver*, *sex*driver* of the second order interaction parameters were significant (p-value<0.05), while *time*result* of the second order interaction parameters was not significant (p-value >0.05). Also, the main terms *result*, *sex* and *driver*

Table 7 - Tests of Partial associations (n=276)

Effect name	Abbreviations	df	Partial Chi-Square	p-value
TIME*RESULT	TR	2	5.768	0.056
SEX*TIME	ST	2	6.589	0.037
TIME*DRIVER	TD	2	18.790	0.000
RESULT*SEX	RS	1	6.107	0.013
RESULT*DRIVER	RD	1	4.333	0.037
SEX*DRIVER	SD	1	5.285	0.022
TIME	T	2	0.152	0.927
RESULT	R	1	20.085	0.000
SEX	S	1	29.891	0.000
DRIVER	D	1	24.734	0.000

were significant (p-value <0.05), while the main term *accident time* was not significant (p-value >0.05).

This study researched which parameters were significant and which model was the best-fit for the data set. There are two techniques for determining the significance of the components in a saturated model: the backward elimination process, and the forward addition process. The backward elimination process was utilized in this study. The backward elimination process begins with all the elements of the saturated model and eliminates the effects one at a time from the highest to the lowest order.

The results of the backward elimination for the selection of the best-fit model can be seen in *Table 8*. In

Table 8 - The results of the backward elimination search for the best-fit model

Step in the modelling process	Deleted effect	Abbreviations	Chi-Square	df	p-value
1	SEX*DRIVER*RESULT*TIME	SDRT	1.355	2	0.508
2	TIME*DRIVER*SEX TIME*DRIVER*RESULT TIME*SEX*RESULT DRIVER*SEX*RESULT	TDS TDR TSR DSR	0.654 1.617 5.804 2.171	2 2 2 1	0.721 0.445 0.055 0.141
3	TIME*DRIVER*RESULT TIME*SEX*RESULT DRIVER*SEX*RESULT	TDR TSR DSR	2.162 6.026 1.751	2 2 1	0.339 0.049 0.186
4	TIME*SEX*RESULT DRIVER*SEX*RESULT TIME*DRIVER	TSR DSR TD	7.018 1.824 19.823	2 1 2	0.030 0.177 0.000
5	TIME*SEX*RESULT TIME*DRIVER DRIVER*SEX DRIVER*RESULT	TSR TD DS DR	5.877 18.610 5.105 4.152	2 2 1 1	0.053 0.000 0.024 0.042
6	TIME*DRIVER DRIVER*SEX DRIVER*RESULT TIME*SEX TIME*RESULT SEX*RESULT	TD DS DR TS TR SR	18.787 5.282 4.331 6.586 5.766 6.104	2 1 1 2 2 1	0.000 0.022 0.037 0.037 0.056 0.013
7	TIME*DRIVER DRIVER*SEX DRIVER*RESULT TIME*SEX SEX*RESULT	TD DS DR TS SR	20.207 5.067 5.749 7.054 6.572	2 1 1 2 1	0.000 0.024 0.016 0.029 0.010
8 Generating class	TIME*DRIVER SEX*DRIVER RESULT*DRIVER TIME*SEX RESULT*SEX	TD SD RD TS RS	17.639	11	0.090

Table 8 the "Deleted effect" is the change in the Chi-Square after the effect is deleted from the model. At each step the effect with the largest significance level for the Likelihood Ratio change is deleted, provided the significance level is larger than 0.05.

In *Table 8* the process begins with the saturated model. The 4-way interaction term is removed but this does not have a significant effect (p-value>0.05). The three-way interaction terms are eliminated one at a time as the next steps (p-value>0.05). Thus, non-significant interaction terms are removed one at a time until all those left are significant. The process then ends and concludes that the best-fit model for this data set has the generating class:

$$\log m_{ijkl} = \mu + \lambda_i^S + \lambda_j^D + \lambda_k^T + \lambda_l^R +$$
$$+ \lambda_{ij}^{SD} + \lambda_{ik}^{ST} + \lambda_{il}^{SR} + \lambda_{jk}^{DT} + \lambda_{jl}^{DR}$$

According to this model, it is observed that *accident time*accident severity* of the second order interaction of the variables *accident severity* (R) and *accident time* (T) is insignificant. The remaining bilateral interaction terms namely *sex*driver's fault and carelessness*, *sex*accident time*, *sex*accident severity*, *accident time*driver's fault and carelessness*, *driver's fault and carelessness*accident severity* have been included in the model. After acceptance of the best model, the estimates of parameters were obtained as follows:

The coefficients (estimates, λ) can be used to estimate the cell frequencies in *Table 9*. Also, these estimates show the dependency of related categories of the variables. These coefficients may be standardized by being divided by their standard errors. Such standardized parameters are identified as "Z-value" by SPSS, because their significance can be evaluated via the standard normal curve. The estimation of the standardized parameter (Z) gives an idea of which categories' relation is most powerful and so the driver's fault and carelessness have the highest value among the main effects., It is understood, namely, that the most important factor for determining of frequencies in the contingency table is "driver's fault and carelessness".

Since the coefficients must sum to zero across categories of a variable, the redundant parameter in *Table 9* is calculated. Accordingly, the parameter estimations for the main effect of driver's fault and carelessness are as follows:

- Driver's fault and carelessness=yes, $\lambda = 1.384$,
- Driver's fault and carelessness=no, $\lambda = -1.384$.

The large λ for driver's fault and carelessness reflects the fact that most of the participants answered "Yes". These main effect (marginal frequencies) coefficients may be quite important for predicting the frequency of a given cell, when the main effects do differ from one another, but they are generally not of great interest otherwise.

Since accident time variable has three categories, the parameter estimations for this variable are as follows:

- accident time=daytime, $\lambda = 0.265$,
- accident time=night time, $\lambda = 0.712$.
- accident time=holiday, $\lambda = -0.977$.

To obtain $\lambda = -0.977$ value is calculated as [0-(0.265+0.712)].

The parameter estimations for the *driver's fault and carelessness*accident time* interaction effects are as follows:

- Driver's fault and carelessness=yes /accident time=daytime is $\lambda = -1.212$
- Driver's fault and carelessness=no / accident time=daytime is $\lambda = +1.212$
- Driver's fault and carelessness=yes / accident time=night time is $\lambda = -1.407$
- Driver's fault and carelessness=no / accident time=night time is $\lambda = +1.407$
- Driver's fault and carelessness=yes / accident time=holiday is $\lambda = +2.619$
- Driver's fault and carelessness=no / accident time=holiday is $\lambda = -2.619$

To obtain $\lambda = +2.619$ value is calculated as [0-(-1.212-1.407)].

Here, "a" is set to zero because it is redundant parameter.

The secondary effective terms evaluated by standardized parameters estimate can be interpreted as follows:

- An accident being due to the driver's fault and carelessness are increased in male drivers.
- An accident being due to the driver's fault and carelessness are increased for holiday time accidents.
- Drivers are prevalently females in holiday time accidents.
- Drivers are prevalently males in daytime and night time accidents.
- Death/injury accidents are accounted for prevalently by females.

5. CONCLUSION

Log-linear models are a type of generalised linear model. The case study demonstrated that log-linear models can help to identify the detailed patterns of interaction between the variables in a contingency table. Log-linear models are commonly used to analyse the relationship between variables in multidimensional tables. The goal of using log-linear modelling procedures is usually to identify the simplest model that fits the data adequately. We preferred to use in this study the log-linear model as it gives value results and also provides information on multiple interactions between various factors which are useful for understanding the overall problem.

Table 9 - The estimation of parameters of the best model

Parameter	Estimate (λ)	Standard Error	Z	p-value	95%Confidence Interval	
					Lower bound	Upper Bound
Constant	1.397	0.316	4.427	0.000	0.778	2.015
D=yes	1.384	0.333	4.159	0.000	0.732	2.036
D=no	0a	.	.	.	.	.
R=injury/death	0.289	0.257	1.128	0.259	-0.214	0.792
R= economic losses	0a	.	.	.	.	.
S=male	0.110	0.341	0.323	0.747	-0.558	0.778
S=female	0a	.	.	.	.	.
T=daytime	0.265	0.349	0.760	0.447	-0.419	0.949
T=night time	0.712	0.326	2.182	0.029	0.072	1.352
T=holiday	0a	.	.	.	.	.
D=yes*R= injury/death	-0.635	0.264	-2.400	0.016	-1.153	-0.116
D=yes*R= economic losses	0a	.	.	.	.	.
D=no*R= injury/death	0a	.	.	.	.	.
D=no*R= economic losses	0a	.	.	.	.	.
D=yes*S=male	0.642	0.286	2.246	0.025	0.082	1.201
D=yes*S=female	0a	.	.	.	.	.
D=no*S=male	0a	.	.	.	.	.
D=no*S=female	0a	.	.	.	.	.
T=daytime*D=yes	-1.212	0.355	-3.412	0.001	-1.909	-0.516
T=daytime*D=no	0a	.	.	.	.	.
T=night time*D=yes	-1.407	0.346	-4.065	0.000	-2.086	-0.729
T=night time*D=no	0a	.	.	.	.	.
T=holiday*D=yes	2.619	.	.	.	.	.
T=holiday*D=no	0a	.	.	.	.	.
R=injury/death*S=male	-0.683	0.266	-2.566	0.010	-1.205	-0.161
R=injury/death*S=female	0a	.	.	.	.	.
R= economic losses *S=male	0a	.	.	.	.	.
R= economic losses *S=female	0a	.	.	.	.	.
T=daytime*S=male	0.877	0.336	2.608	0.009	0.218	1.536
T=daytime*S=female	0a	.	.	.	.	.
T=night time* S=male	0.474	0.320	1.481	0.139	-0.154	1.102
T=night time* S=female	0a	.	.	.	.	.
T=holiday* S=male	-1.351	.	.	.	.	.
T=holiday* S=female	0a	.	.	.	.	.

This study has assessed the relationship between driver's fault and carelessness and the traffic variables such as sex, accident severity, and accident time. In this study, 1,325 participants, in four different districts of Ankara, were asked in a questionnaire if they had been involved in any traffic accidents in the last one-year period. Among the people participating in this questionnaire, only 448 of them stated that they had been involved in an accident in the year 2008. As drivers, 276 out of 448 people, namely 61.6%, had had traffic accidents. This ratio of being involved in a traffic accident is actually considerably high in Ankara. As drivers having traffic accidents, 179 of them, namely 64.9%, involved driver's fault and carelessness. In addition to this 101 drivers, namely 36.6%, had injury/death as result of the accident. These figures are indicators of how significant the traffic problem is in Turkey. It shows that many years of efforts and education programmes have not really helped in solving the problem, even though Ankara is a city with a high level of education. It has been noticed in this study that the factors sex, accident time, driver's fault and carelessness, are significantly associated with accident severity. It is evident that the accident severity is highly related to sex, accident time and driver's fault and carelessness. In terms of the accident severity variable, approximately 36.6% of the drivers are classified as injury/death. The female drivers have a higher proportion (48.4%) of injury/death while the proportion of male drivers is slightly lower (30.6%). The group driver's fault and carelessness tends to have a higher proportion of economic losses (69.2%). The night time tends to have a higher injury/death proportion (47.4%), while the daytime proportion is slightly lower (28.9%).

It is noticed that there is a 0.764 greater probability of injury/death for males than for females. The relative risk of injury/death is 0.769 for the variable of driver's fault and carelessness. Thus, there is a 0.769 greater probability of injury/death for driver's fault and carelessness than without driver's fault and carelessness. Also, there are nearly twofold greater odds of injury/death for night time than for holiday time.

This study researched which parameters were significant and which model was the best-fit for the data set. It has concluded that the best-fit model for this data set has the generating class. According to this model, it is observed that *sex*driver's fault and carelessness*, *sex*accident time*, *sex*accident severity*, *accident time*driver's fault and carelessness*, *driver's fault and carelessness*accident severity* have been included in the model. After having accepted the best model, the estimates of parameters were obtained.

Driver's fault and carelessness factors are the most important factors affecting the occurrence of traffic accidents. Driver's fault and carelessness increase in holiday time accidents. Drivers in daylight accidents are more likely to be male. Death/injury accidents are prevalently accounted for by females. Finally, it can be mentioned that valuable results were obtained with log-linear models. We have concluded that driver's fault and carelessness mostly increased in female drivers and during holiday time.

This study has shown that driver's fault and carelessness factors affecting the occurrence of traffic accidents are the most important factors in spite of there being a large number of measures (money penalties, police controls, radar controls...) for the prevention of accidents. It has also been seen that in the accidents caused by driver's fault and carelessness, death and injuries are important outcomes.

The male participants were involved more than the female participants in the last one-year period. The most common reason for an accident stated by the participants of the questionnaire, and the ones involved in an accident in the last one-year period, was "driver's fault and carelessness". Identification of these risk factors of accident severity will provide information to injury/death preventive programmers. Most of the participants suggested the view that "traffic education programmes should be more commonly applied" as the precautions and efforts needed to be implemented for preventing these accidents.

HÜLYA OLMUŞ
E-mail: hulya@gazi.edu.tr
SEMRA ERBAŞ
E-mail: serbas@gazi.edu.tr
Gazi Üniversitesi, Fen Fakültesi
Istatistik Bölümü
06500 Teknikokullar-Ankara, Türkiye

ÖZET

LOG-LINEER MODELLER KULLANILARAK SÜRÜCÜLERIN NEDEN OLDUGU TRAFIK KAZALARININ ANALIZI

Log-lineer modelleme degişik özelliklerin ortaya çikmasinda relatif frekansin temelini oluşturan faktörleri belirlemek için bir yöntem olarak geliştirilmiştir. Bu çalişmanin amaci, cinsiyet, kaza ciddiyeti ve kaza zamani gibi trafik degişkenleri ve sürücü hata ve dikkatsizligi degişkenleri arasindaki ilişkileri tahmin etmek için log-lineer model kullanilarak bir modelleme sunmaktir. Bu çalişma, Türkiye'nin başkenti olan Ankara'da 4 farkli ilçede yürütüldü. Çalişma için 1,325 birey seçildi ve bu bireylere kazada olup olmadiklari soruldu. Bu bireylerin 448'i kazada yer aldiklarini ifade etti. 448 insandan 276'si yani %61.6'si sürücü olarak trafik kazasinda yer almiştir. Cinsiyet, sürücünün hata ve dikkatsizligi, kaza ciddiyeti, kaza zamani degişkenlerinin yer aldigi veri, anket çalişmasi süresince toplandi. Ayrintili bilgiler, bu bilgiye dayali oluşturuldu. Analizler, bu degişkenlere ilişkin en iyi modelin log-lineer model oldugunu gösterdi. Ayrica, bu degişkenler arasindaki odds oranlari, kaza ciddiyeti ile faktörler arasindaki ilişkiler, degişik faktörlerin katkilari ve bu degişkenler arasindaki çoklu etkileşimler degerlendirildi. Elde edilen sonuçlar, trafik kazalarinda istenmeyen sonuçlari engellemek için önemli bilgi saglar.

ANAHTAR KELIMELER

Log-lineer model, trafik kazalari, odds orani, olabilirlik-oran test istatistigi

LITERATURE

[1] **Abdel-Aty, M.**, **Chen, C.L.**, **Schott, J.R.**: *An assessment of the effect of driver age on traffic accident involvement using log-linear models*. Accident Analysis and Prevention 30(6), 1998, 851-861

[2] **Agresti, A.**: *Categorical Data Analysis*. John Wiley, New York, 1990

[3] **Decarlo, E. T.**, **Laczniak, R.N.**, **Azevedo K.A.**, **Ramaswami, S.N.**: *On the Log-Linear Analysis of multiple response data*. Marketing Letters 11(4), 2000, 349-361

[4] **Iacobucci, D.**, **Ann, L. McGill.**: *Analysis of Attribution Data: Theory Testing and Effects Estimation*. Journal of Personality and Social Psychology. 59(3), 1990, 426-441

[5] **Jang, T.Y.**: *Analysis on reckless driving behavior by log-linear model*, KSCE Journal of Civil Engineering, 10 (4), 2006, 297-303

[6] **Kim, K., Nitz, L., Richardson, J., Li, L.**: *Analyzing the relationship between crash types and injuries in motor vehicle collisions in Hawaii*. Transportation Research Record.1467, 1995a, 9-13

[7] **Kim, K., Nitz, L., Richardson, J., Li, L.**: *Personal and behavioral predictors of automobile crash and injury severity*, Accident Analysis and Prevention, 27, 4, 1995b, 469-481

[8] **Lawal, B.**: *Categorical Data Analysis with SAS and SPSS Applications*, Lawrence Erlbaum Associates, Inc., London, (2003)

[9] **Lourens, P. F., Vissers, J. A. M. M., Jessurun, M.**: *Annual mileage, driving violations and accident involvement in relation to drivers' sex, age and level of education*. Accident Analysis and Prevention, 31 (1), 1999, 593-597

[10] **Richardson, J., Kim, K., Li, L., Nitz, L.**: *Patterns of motor vehicle crash involvement by driver age and sex in Hawaii*. Journal of Safety Research. 27(2), 1996, 117-125

[11] **Upton, J.G.**: *The Analysis of Cross-tabulated Data*. John Wiley and Sons, New York, 1977