PERGAMON

Accident Analysis and Prevention 34 (2002) 729–741

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Using logistic regression to estimate the influence of accident factors on accident severity

Ali S. Al-Ghamdi *

College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia

Received 13 June 2000; received in revised form 28 May 2001; accepted 1 June 2001

Abstract

Logistic regression was applied to accident-related data collected from traffic police records in order to examine the contribution of several variables to accident severity. A total of 560 subjects involved in serious accidents were sampled. Accident severity (the dependent variable) in this study is a dichotomous variable with two categories, fatal and non-fatal. Therefore, each of the subjects sampled was classified as being in either a fatal or non-fatal accident. Because of the binary nature of this dependent variable, a logistic regression approach was found suitable. Of nine independent variables obtained from police accident reports, two were found most significantly associated with accident severity, namely, location and cause of accident. A statistical interpretation is given of the model-developed estimates in terms of the odds ratio concept. The findings show that logistic regression as used in this research is a promising tool in providing meaningful interpretations that can be used for future safety improvements in Riyadh.

Keywords: Logistic regression; Accident severity

1. Introduction

Accident severity is of special concern to researchers in traffic safety since this research is aimed not only at prevention of accidents but also at reduction of their severity. One way to accomplish the latter is to identify the most probable factors that affect accident severity. This study aims at examining not all factors, but some believed to have a higher potential for serious injury or death, such as accident location, type, and time; collision type; and age and nationality of the driver at fault, his license status, and vehicle type. Other factors were not examined because of substantial limitations in the data obtained from accident reports. Logistic regression was used in this study to estimate the effect of the statistically significant factors on accident severity. Logistic regression and other related categorical-data regression methods have often been used to assess risk factors for various diseases. However, logistic regression has been used as well in transportation studies. A brief literature review follows of the use of this type of regression in traffic safety research.

2. Literature review

Regression methods have become an integral component of any data analysis concerned with the relationship between a response variable and one or more explanatory variables. The most common regression method is conventional regression analysis (CRA), either linear or nonlinear, when the response variable is continuous (iid). However, when the outcome (the response variable) is discrete, CRA is not appropriate. Among several reasons, the following two are the most significant:

1. The response variable in CRA must be continuous, and
2. The response variable in CRA can take nonnegative values.

These two primary assumptions are not satisfied when the response variable is categorical.

* Tel.: +966-1-467-7019; fax: +966-1-467-4254.
E-mail address: asghamdi@ksu.edu.sa (A.S. Al-Ghamdi).

0001-4575/02/$ - see front matter
PII: S0001-4575(01)00073-2

Jovanis and Chang (1986) found a number of problems with the use of linear regression in their study applying Poisson regression as a means to predict accidents. For example, they discovered that as vehicle-kilometers traveled increases, so does the variance of the accident frequency. Thus, this analysis violates the homoscedasticity assumption of linear regression.

In a well-summarized review of models predicting accident frequency, Milton and Mannering (1997) state: "The use of linear regression models is inappropriate for making probabilistic statements about the occurrences of vehicle accidents on the road." They showed that the negative binomial regression is a powerful predictive tool and one that should be increasingly applied in future accident frequency studies.

Kim et al. (1996) developed a logistic model and used it to explain the likelihood of motorists being at fault in collisions with cyclists. Covariates that increase the likelihood of motorist fault include motorist age, cyclist age (squared), cyclist alcohol use, cyclists making turning actions, and rural locations.

Kim et al. (1994) attempted to explain the relationship between types of crashes and injuries sustained in motor vehicle accidents. By using techniques of categorical data analysis and comprehensive data on crashes in Hawaii during 1990, a model was built to relate the type of crash (e.g. rollover, head-on, sideswipe, rear-end, etc.) to a KABCO injury scale. They also developed an 'odds multiplier' that enabled comparison according to crash type of the odds of particular levels of injury relative to noninjury. The effects of seatbelt use on injury level were also examined, and interactions among belt use, crash type, and injury level were considered. They discussed how log-linear analysis, logit modeling, and estimation of 'odds multipliers' may contribute to traffic safety research.

Kim et al. (1995) built a structural model relating driver characteristics and behavior to type of crash and injury severity. They explained that the structural model helps to clarify the role of driver characteristics and behavior in the causal sequence leading to more severe injuries. They estimated the effects of various factors in terms of odds multipliers — that is, how much does each factor increase or decrease the odds of more severe crash types and injuries.

Nassar et al. (1997) developed an integrated accident risk model (ARM) for policy decisions using risk factors affecting both accident occurrences on road sections and severity of injury to occupants involved in the accidents. Using negative binomial regression and a sequential binary logit formulation, they developed models that are practical and easy to use. Mercier et al. (1997) used logistic regression to determine whether either age or gender (or both) was a factor influencing severity of injuries suffered in head-on automobile collisions on rural highways.

Logistic regression was also used by Hilakivi et al. (1989) in predicting automobile accidents of young drivers. They examined the predictive values of the Cattel 16-factor personality test on the occurrence of automobile accidents among conscripts during 11-month military service in a transportation section of the Finnish Defense Forces.

James and Kim (1996) developed a logistic regression model to describe the use of child safety seats for children involved in crashes in Hawaii from 1986 through 1991. The model reveals that children riding in automobiles are less likely to be restrained, drivers who use seat belts are far more likely to restrain their children, and 1- and 2-year-olds are less likely to be restrained.

3. Theoretical background of logistic regression

It is important to understand that the goal of an analysis using logistic regression is the same as that of any model-building technique used in statistics: to find the best fit and the most parsimonious one. What distinguishes a logistic regression model from a linear regression model is the response variable. In the logistic regression model, the response variable is binary or dichotomous. The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression analysis. In any regression analysis the key quantity is the mean value of the response variable given the values of the independent variable:

$$E(Y/x) = \beta_0 + \beta_1 x$$

where Y denotes the response variable, x denotes a value of the independent variable, and the β_i-values denote the model parameters. The quantity is called the conditional mean or the expected value of Y given the value of x. Many distribution functions have been proposed for use in the analysis of a dichotomous response variable (Hosmer and Lemeshow, 1989; Agresti, 1984; Feinberg, 1980). The specific form of the logistic regression model is

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (1)$$

where, to simplify the notation, $\pi(x) = E(Y/x)$. The transformation of the $\pi(x)$ logistic function is known as the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x \quad (2)$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from minus infinity to plus infinity, depending on the range of x.

Hosmer and Lemeshow (1989) summarize the main features in a regression analysis when the response variable is dichotomous:

1. The conditional mean of the regression equation must be formulated to be bounded between zero and 1 (Eq. (1) satisfies this constraint).
2. The binomial, not the normal, distribution describes the distribution of the errors and will be the statistical distribution upon which the analysis is based.
3. The principles that guide an analysis using linear regression will also apply for logistic regression.

In linear regression the method used most often for estimating unknown parameters is least squares, in which the parameter values are chosen to minimize the sum of squared deviations of the observed values of Y from the modeled values. Under the assumptions for linear regression, the method of least squares yields estimators with a number of desirable statistical properties. Unfortunately, when the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these same properties. The general method of estimation that leads to the least squares function under the linear regression model (when the error is normally distributed) is called maximum likelihood. This method provides the foundation for estimating the parameters of a logistic regression model. A brief review of fitting the logistic regression model is given below. Further details may be found elsewhere (Hosmer and Lemeshow, 1989).

If Y is coded as zero or 1 (a binary variable), the expression $\pi(x)$ given in Eq. (1) provides the conditional probability that Y is equal to 1 given x, denoted as $P(Y=1/x)$. It follows that the quantity $1-\pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(Y=0/x)$. Thus, for those pairs (x_i, y_i) where $y_i=1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i=0$, the contribution to the likelihood function is $1-\pi(x_i)$, where the quantity $\pi(x_i)$ denotes the values of $\pi(x)$ computed at x_i. A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the term

$$\zeta(x_i)=\pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i}$$

Since x_i-values are assumed to be independent, the product for the terms given in the foregoing equation gives the likelihood function as follows:

$$l(\beta)=\prod_{i=1}^{n}\zeta(x_i) \tag{3}$$

It is easier mathematically to work with the log of Eq. (3), which gives the log likelihood expression:

$$L(\beta)=\ln\,[l(\beta)]$$

$$=\sum_{i=1}^{n}\{y_i\ln\,[\pi(x_i)]+(1-y_i)\ln\,[1-\pi(x_i)]\} \tag{3.1}$$

Maximizing the above function with respect to β and setting the resulting expressions equal to zero will produce the following values of β:

$$\sum_{i=1}^{n}[y_i-\pi(x_i)]=0 \tag{4}$$

$$\sum_{i=1}^{n}x_i[y_i-\pi(x_i)]=0 \tag{5}$$

These expressions are called likelihood equations. An interesting consequence of Eq. (4) is

$$\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\hat{\pi}(x_i)$$

That is, the sum of the observed values of y is equal to the sum of the expected (predicted) values. This property is especially useful in assessing the fit of the model (Hosmer and Lemeshow, 1989).

After the coefficients are estimated, the significance of the variables in the model is assessed. If y_i denotes the observed value and $\hat{y}_i$ denotes the predicted value for the ith individual under the model, the statistic used in the linear regression is

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

The change in the values of SSE is due to the regression source of variability, denoted SSR:

$$SSR=\text{Total Sum of Squares (SS)}$$

$$-\text{Sum of Squares of Error term (SSE)}$$

$$=\left[\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\right]-\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right]$$

where $\bar{y}$ is the mean of the response variable. Thus, in linear regression, interest focuses on the size of R. A large value suggests that the independent variable is important, whereas a small value suggests that the independent variable is not useful in explaining the variability in the response variable.

The principle in logistic regression is the same. That is, observed values of the response variable should be compared with the predicted values obtained from models with and without the variable in question. In logistic regression this comparison is based on the log likelihood function defined in Eq. (3.1). Defining the saturation model as one that contains as many parameters as there are data points, the current model is the one that contains only the variable under question. The likelihood ratio is as follows:

$$D = -2\ln\left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}}\right] \quad (6)$$

Using Eqs. (3.1) and (6), the following test statistic can be obtained:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \quad (7)$$

where $\hat{\pi}_i = \hat{\pi}(x_i)$.

The statistic D in Eq. (7), for the purpose of this study, is called the deviance, and it plays an essential role in some approaches to the assessment of goodness of fit. The deviance for logistic regression plays the same role that the residual sum of squares plays in linear regression (i.e. it is identically equal to SSE).

For the purpose of assessing the significance of an independent variable, the value of D should be compared with and without the independent variable in the model. The change in D due to inclusion of the independent variable in the model is obtained as follows:

$$G = D(\text{for the model without the variable}) - D(\text{for the model with the variable})$$

This statistic plays the same role in logistic regression as does the numerator of the partial F-test in linear regression. Because the likelihood of the saturated model is common to both values of D being the difference to compute G, this likelihood ratio can be expressed as

$$G = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right] \quad (8)$$

It is not appropriate here to derive the mathematical expression of the statistic G. Yet it should be said that under the null hypothesis, β_1 is equal to zero, G will follow a χ^2 distribution with one degree of freedom. Another test statistic, similar to G for the purpose used in this study, is known as the Wald statistic (W), which follows a standard normal distribution under the null hypothesis that $\beta_1 = 0$. This statistic is computed by dividing the estimated value of the parameter by its standard error:

$$W = \frac{\hat{\beta}_1}{S\hat{E}(\hat{\beta}_1)} \quad (9)$$

It should be mentioned that the Wald test behaved in an aberrant manner, often failing to reject the null hypothesis when the coefficient was significant, and hence the likelihood ratio test should be used in suspicious cases.

4. Model description

The dependent variable in this research, *ACCIDENT*, is of the dichotomous type and stands for accident severity. It should be mentioned that the definition of injury in this study does not overlap with the definition of fatality since the first includes those who were involved in accidents and left the hospital within 6 months after treatment. Each accident in the sampled data was categorized as either non-fatal or fatal. The logistic model used is

$$P(\text{non-fatal accident}) = \pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (10)$$

and thus

$$P(\text{fatal accident}) = 1 - P(\text{injury accident}) = 1 - \pi(x) = \frac{1}{1+e^{g(x)}}$$

where $g(x)$ stands for the function of the independent variables:

$$g(x) = \beta_o + \beta_1 x_1 + \beta_1 x_2 + \cdots + \beta_n x_n$$

Logistic regression determines the coefficients that make the observed outcome (non-fatal or fatal accident) most likely using the maximum-likelihood technique. The independent variables could be continuous or dichotomous, as will be discussed in the next section. For the latter, there should be special coding with the use of dummy variables. These dummy variables should be defined in a manner consistent with the generalized linear interactive modeling (GLIM) software used in this study (GLIM, 1987). The Wald tests, together with the deviance, will be used as criteria to include or remove independent variables from the model. The GLIM software has built-in routines to obtain deviance and estimates of the model parameters.

5. Data description

The data set used in this study was derived from a sample of 560 subjects involved in serious accidents reported in traffic police records in Riyadh, the capital of Saudi Arabia. Only accidents occurring on urban roads in Riyadh were examined. Unfortunately, police reports at accident sites do not describe injuries in much detail because of the lack of police qualifications and training as well as facilities needed to perform complex examinations. Also, medical reports are hard to obtain because police accident data and medical data are not kept together (Al-Ghamdi, 1996). Consequently, it was impossible for this study to obtain details on the degree of severity of the accidents. All that can be learned from the police records is that the accident is a property damage only (PDO) accident, injury accident (no injury classification is available), or fatal accident. The subjects were selected in a systematic random process from all accident records filed for the period from August 1997 to November 1998. The data search was done manually because of the lack of com-

puterization. Only injury (non-fatal) and fatal accident records were considered for the purpose of this study. Since the study goal was to identify the factors that might affect the severity of the accident (i.e. whether it was a fatal or non-fatal accident), 10 variables were summarized from the data. The description and levels of these variables are given in Table 1. Fig. 1 shows the age distribution for drivers in the data set.

The response variable is Variable 1, namely, *ACCIDENT*, which is binary (dichotomous) in nature. Two levels for *ACCIDENT* exist: 0 if the accident results in at least one injury but no fatality (within 6 months after the accident), and 1 if there is at least one fatality resulting from the accident. For the explanatory variables (independent variables), age is the only continuous variable; the others are categorical. Since some of the categorical variables have several levels, identified as 1, 2, 3, and so forth, a collection of design variables (or dummy variables) was needed to represent the data and match the format of GLIM (1987), the software used in this study.

Table 1
Description of the study variables

Number	Description	Codes/values	Abbreviation
1	Accident	0 = Non-fatal 1 = Fatal	*ACCIDENT*
2	Location	1 = Intersection 2 = Non-intersection	*LOC*
3	Accident type	1 = With vehicle(s) 2 = Fixed-object 3 = Over-turn 4 = Pedestrian	*ATYP*
4	Collision type	1 = Right-angle 2 = Sideswipe 3 = Rear-end 4 = Front 5 = Unknown	*CTYP*
5	Accident time	1 = Day 2 = Night	*TIME*
6	Accident cause	1 = Speed 2 = Run red light 3 = Follow too close 4 = Wrong way 5 = Failure to yield 6 = Other	*CAUS*
7	Driver age at fault	Years	*AGE*
8	Nationality	1 = Saudi 2 = Non-Saudi	*NAT*
9	Vehicle type	1 = Small passenger car 2 = Large passenger car 3 = Pick-up truck 4 = Taxi 5 = Other	*VEH*
10	License status	1 = Yes (valid) 2 = Expired 3 = No (no license)	*LIC*

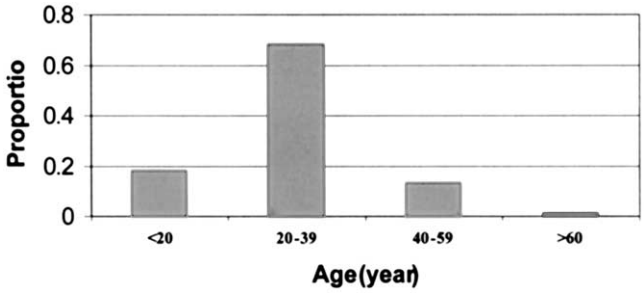

Fig. 1. Distribution of drivers by age.

One possible way of coding the dummy variables is to have $k-1$ design variables for the k levels of the nominal scale of that variable. An example of this coding is given in Table 2 for the variable Accident type (*ATYP*), which has four levels, and hence has three design variables. When the respondent is 'With vehicle(s)', the three design variables, D_1, D_2, and D_3, would all be set to equal zero; when the respondent is 'Fixed object,' D_1 would be set equal to 1 whereas D_2 and D_3 would still equal 0; and so forth for the other respondents. This coding scheme was used for the rest of the categorical variables. It should be noted that GLIM has the capability to do this coding automatically once the levels of the variables have been identified by the end user. Other software packages might use different strategies for coding design variables. It is important to understand the coding strategy used in the software package in order to conduct hypothesis testing on the variables as well as to interpret their estimates.

6. Reduction of design variables

As can be seen from Table 1, some of the categorical variables have several levels, so several design variables are needed for each. Generally speaking, it is more

Table 2
The design variables for Accident type

ATYP	Design variable		
	D_1	D_2	D_3
With vehicle(s)	0	0	0
Fixed-object	1	0	0
Overturn	0	1	0
Pedestrian	0	0	1

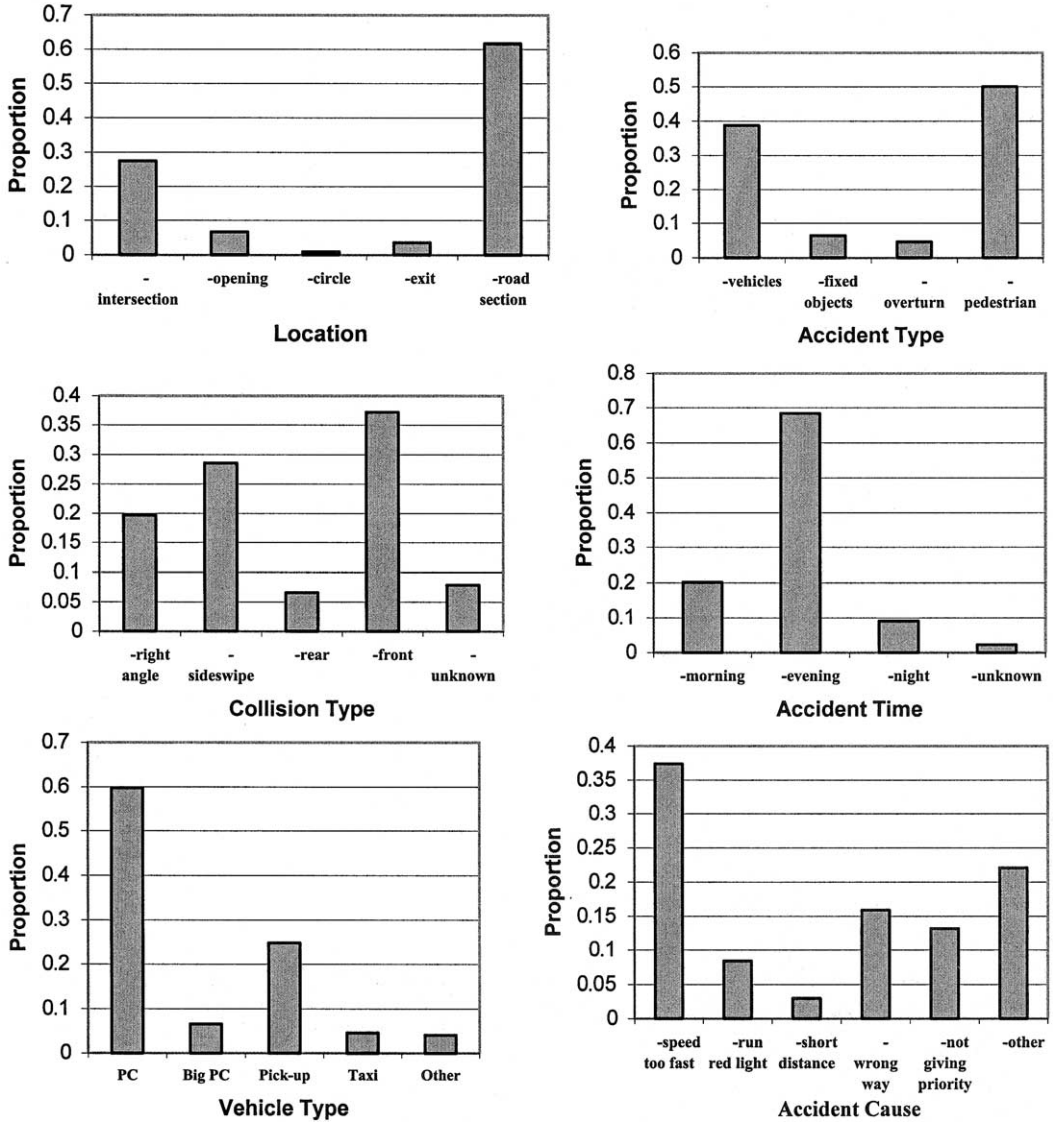

Fig. 2. Study variables.

convenient to have as few design variables as possible in order to simplify the model interpretation. In other words, the more design variables the model includes, the more difficult the interpretation becomes. Thus, an attempt was made in the early stages of this study to reduce the number of design variables. However, care is needed in doing so to guarantee that the model will not lose significant information.

Looking at the proportion of the levels for the study variables (Fig. 2), one can see that some levels can be neglected because of their small proportion. However, such a cursory investigation is not enough to decide which levels can be neglected or at least merged with other levels. Thus, the hypothesis testing technique for proportions was used in this study to decide whether the number of levels for a design variable could be reduced. The following typical test was used:

$$H_o: p_i = 0$$

$$H_a: p_i \neq 0$$

where p_i is the proportion of class i (level i) within the designated design variable.

For example, the design variables for *ATYP* were reduced from three (four levels) to two (three levels) after it was shown that the proportion of fixed-object and overturn accidents was not statistically significant at the 5% level using the foregoing hypothesis. Table 3 summarizes the hypothesis testing results for all categorical variables in the study, and Table 4 shows the number of design variables after reduction.

The study variables were now ready to use in the model development stage, as discussed in the next section.

7. Development of logistic model

The backward selection process of logistic regression was followed. First, all the variables with no interactions (referred to here as the saturated model; Fig. 3) were tested on the basis of the deviance and the Wald (W) statistic as defined in Eqs. (7) and (9), respectively. The goal was to eliminate, at the beginning, those variables that were not significant and then continue with testing interaction effects with only significant variables. Table 5 presents the results from fitting all the explanatory variables simultaneously. From the W-values (Table 5), it appears that the variables LOC, $CAUS$, AGE, NAT, and LIC show some significant effect (AGE, NAT, and LIC are about significant); however, further testing using deviance is needed. Because of the multiple degrees of freedom, one must be careful in the use of the Wald (W) statistic to assess the significance of the coefficients. For example, the variable $CAUS$ has five levels, but only two of the levels $CAUS(4)$ and $CAUS(5)$ were found to be statistically significant at the 0.05 level (Table 5). In this case, the decision to include this variable should be made using the likelihood ratio test. That is, the change in deviance

Table 3
Hypothesis testing for proportions

Description	x	n	P-value	95% confidence limits	
				Lower	Upper
Distribution by location					
Intersection	159	579	0.275	0.2	0.3
Opening*	38	579	0.066	0	0.1
Circle*	5	579	0.009	0	0
Exit*	20	579	0.035	0	0
Road section	357	579	0.617	0.6	0.7
Distribution by accident type					
Vehicles	235	605	0.388	0.4	0.4
Fixed objects*	39	605	0.064	0	0.1
Overturn*	28	605	0.046	0	0.1
Pedestrian	303	605	0.501	0.5	0.5
Distribution by collision type					
Right angle	119	605	0.197	0.2	0.2
Sideswipe	173	605	0.286	0.3	0.3
Rear*	40	605	0.066	0	0.1
Front	225	605	0.372	0.3	0.4
Unknown	48	605	0.079	0.1	0.1
Distribution by time					
Morning	122	605	0.202	0.2	0.2
Evening	414	605	0.684	0.7	0.7
Night	55	605	0.091	0.1	0.1
Unknown*	14	605	0.023	0	0
Distribution by accident cause					
Speed too fast	226	605	0.374	0.3	0.4
Run red light	51	605	0.084	0.1	0.1
Short distance*	18	605	0.03	0	0
Wrong way	96	605	0.159	0.1	0.2
Not giving priority	80	605	0.132	0.1	0.2
Other	134	605	0.221	0.2	0.2
By vehicle type					
PC	335	560	0.598	0.6	0.6
Big PC*	37	560	0.066	0	0.1
Pick-up	139	560	0.248	0.2	0.3
Taxi*	26	560	0.046	0	0.1
Other*	23	560	0.041	0	0.1
By Licensing status					
Yes	370	560	0.661	0.6	0.7
Expired*	37	560	0.066	0	0.1
No	153	560	0.273	0.2	0.3

* Statistically insignificant at 5% level (the 95% confidence limits include zero).

Table 4
Number of design variables after reduction

Categorical variable	Before reduction		After reduction	
	Levels	Design variables	Levels	Design variables
Location[a]	5	4	2	1
Accident type[a]	4	3	3	2
Collision type[a]	5	4	4	3
Accident time	2	1	2	1
Accident cause[a]	6	5	5	4
Nationality	2	1	2	1
Vehicle type[a]	5	4	2	1
License status[a]	3	2	2	1

[a] Variables experience reduction.

for the model should be assessed both with the variable and without it.

Removal of *LIC* from the model did not produce much change in the deviance, and thus it is not significant at the 0.05 level ($P = 0.093$) as shown in Table 6. This finding indicates that *LIC* is not adding useful information to the variability in the response variable and should be removed. Similarly, the variables *VEH*, *TIME*, *CTYP*, *ATYP*, *AGE*, and *NAT* do not show any major changes in deviance, and accordingly they were dropped from the model. On the other hand, the variables *LOC* and *CAUS* are found to be statistically significant at the 0.05 level (Table 6).

Therefore, the backward selection process identified two variables (*LOC* and *CAUS*) as being significantly related to accident severity. These two variables were then subjected to further analysis, as will be discussed shortly. Before that analysis, it might have been thought that accident type (*ATYP*) and collision type (*CTYP*) would have had a significant effect on accident severity, yet that was not the case in this study since they failed to meet the desired significance level (0.05). However, it might be argued that these two variables are implied in the two significant variables in the model, namely, *LOC* and *CAUS*. For example, since it is known that serious accidents occur at intersections, right-angle collisions would be the most likely type caused by running a red light (note that right-angle collision type and run-red-light accident cause have significant proportions in Table 3). Right-angle collisions caused by running a red light are a common problem in Saudi Arabia (Official Statistics, 1997). Accordingly, the presence of *LOC* and *CAUS* in the model would imply *CTYP*. In the same context, an accident occurring along a roadway section (non-intersection location) would imply a multiple-vehicle, fixed-object, or pedestrian accident (*ATYP*).

```
[i] ? $fit +lic$
 [o] scaled deviance = 491.54 (change =  -2.83) at cycle  4
 [o]         d.f. = 545   (change =  -1  )
 [o]
 [i] ? $dis e d$
 [o]       estimate      s.e.    parameter
 [o]   1    -2.662     0.6416     1
 [o]   2     0.9050    0.3520     LOC(2)
 [o]   3     0.02367   0.3690     ATYP(2)
 [o]   4    -0.3032    0.4490     ATYP(3)
 [o]   5    -0.3218    0.3904     CTYP(2)
 [o]   6    -0.02615   0.3060     CTYP(3)
 [o]   7     0.1194    0.4545     CTYP(4)
 [o]   8    -0.01806   0.3917     TIME(2)
 [o]   9    -0.3071    0.5849     CAUS(2)
 [o]  10     0.1447    0.3140     CAUS(3)
 [o]  11    -0.9131    0.4635     CAUS(4)
 [o]  12    -0.7727    0.3217     CAUS(5)
 [o]  13     0.01883   0.01232    AGE
 [o]  14     0.3469    0.2739     NAT(2)
 [o]  15    -0.2871    0.2406     VEH(2)
 [o]  16     0.4257    0.2514     LIC(2)
```

Fig. 3. A GLIM output for the saturated model.

7.1. Interaction and confounding effects

The two variables found to be statistically significant in the current study (i.e. *LOC* and *CAUS*) were investigated further with the possible term of interaction. The process is to add each interaction term to the full model (i.e. the model with the two significant terms). If the added term is significant, the change in deviance between the full model and the model with the added term (interaction) should be large enough to be statistically significant at the 0.05 level. The interaction was found to be statistically insignificant ($P = 0.265$), as presented in Table 7, and hence a confounding effect does not exist.

7.2. Age effect

Understanding and quantifying the relationship between driver characteristics, particularly age and accident risk, has long been a high priority of accident-related research. In addition, other studies (Hilakivi et al., 1989; Mercier et al., 1997) have shown that young drivers as well as older drivers are more at risk of being involved in serious accidents. Numerous research studies have attempted to examine the complex relationship between driver characteristics and accident risk. Drivers' risk-taking behavior is often defined in terms of several variables, one of which is age. Manner-

Table 5
Estimated coefficients, estimated standard errors, and Wald statistic for the model variables

Variable	Estimated coefficient	Estimated standard error	Wald statistic (W)	P-value
$LOC(2)$	0.905	0.352	2.57	0.005*
$ATYP(2)$	0.02367	0.369	0.06	0.48
$ATYP(3)$	−0.3032	0.449	−0.68	0.75
$CTYP(2)$	−0.3218	0.3904	−0.82	0.79
$CTYP(3)$	−0.02615	0.306	−0.09	0.54
$CTYP(4)$	0.1194	0.4545	0.26	0.40
$TIME(2)$	−0.01806	0.3917	−0.05	0.52
$CAUS(2)$	−0.3071	0.5849	−0.53	0.30
$CAUS(3)$	0.1447	0.314	0.46	0.32
$CAUS(4)$	−0.9131	0.4635	−1.97	0.024*
$CAUS(5)$	−0.7727	0.3217	−2.4	0.01*
AGE	0.01883	0.01232	1.53	0.06
$NAT(2)$	0.3469	0.2739	1.27	0.10
$VEH(2)$	−0.2871	0.2406	−1.19	0.79
$LIC(2)$	0.4257	0.2514	1.69	0.045*

* Statistically significant at 5% level.

ing (1992) indicates that age itself is really being used as a surrogate for drivers' risk-taking behavior. Some researchers also indicate that age relates nonlinearly to the response variable (Mercier et al., 1997; Hosmer and Lemeshow, 1989). They suggest that a quadratic expression be used.

The problem with the age variable in this study appears from the unexpected positive effect shown in the parameter estimate in Table 5. It was expected that the older the driver, the less the accident risk. Safety research in Saudi Arabia has always indicated that age is a primary factor in risk-taking behavior (Official Statistics, 1997; Al-Ghamdi, 1996). Young drivers are involved in about one-fifth of the accidents nationwide (Official Statistics, 1997). Therefore, the author decided to investigate the age factor more closely, even though it had been shown from the analysis in this study that age was not statistically significant. The model has shown so far that age, in a linear relation with the dependent variable, is not statistically significant. Thus, the possible quadratic form was tested as suggested in past research. That is, age-squared (as a quadratic effect) entered the model with the two significant variables (LOC and $CAUS$). The result showed that the quadratic main effect of age was not statistically significant either ($P = 0.52$, Table 7).

8. Logit model

According to the previous analysis, the logit model with the significant variables is as follows:

$$\begin{aligned}\hat{g}(x) &= -2.029 + 0.9697LOC(2) - 0.3558CAUS(2) \\ &\quad + 0.2130CAUS(3) - 0.8971CAUS(4) \\ &\quad - 0.6705CAUS(5)\end{aligned}$$

Hence the logistic regression model developed in this study is

$$\pi(x) = \frac{e^{-2.029 + 0.9697LOC(2) - 0.3558CAUS(2) + 0.2130CAUS(3) - 0.8971CAUS(4) - 0.6705CAUS(5)}}{1 + e^{-2.029 + 0.9697LOC(2) - 0.3558CAUS(2) + 0.2130CAUS(3) - 0.8971CAUS(4) - 0.6705CAUS(5)}}$$

Once the model has been fit, the process of assessment of the model begins. Several tests, including Pearson χ^2 and deviance, the Wald statistic, and the Hosmer–Lemeshow tests, can be used to determine how effective the model is in describing the response variable, or its goodness of fit. These tests resulted in a χ^2 criterion to make the decision on the model fit. A very good source for the theory of such tests is, for example, Hosmer and Lemeshow (1989). The validity of the model in this study was first checked by examining the statistical level of significance for its coefficients using deviance and the Wald statistic, as discussed earlier.

Graphical assessment of the fit to the logistic model developed in the study also shows that the model appears to fit the data reasonably, as shown in Figs. 4 and 5. Fig. 4 shows the plot of Pearson residuals, in which no trend can be detected. Fig. 5 shows Hi-Leverage points (outliers) in which very small points appear to be outliers [less than 4% of the data set; compare $PRES$ with 1.96 (z-value at the 5% level of significance)]. That is, 95% of the points in this plot lie between -0.5 and 1.9.

9. Model interpretation

Interpretation of any fitted model requires the ability to draw practical inferences from the estimated coefficients. The estimated coefficients for the independent variables represent the slope or rate of change of the dependent variable per unit of change in the independent variable. Thus, interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable (i.e. the link function; McCullagh and Nelder, 1982) and appropriately defining the unit change for the independent variable. In the logistic regression model, the link function is the logit transformation (Eq. (2)). The slope coefficient in this model represents the change in the logit for a change of one unit in the independent variable x. Proper interpretation of the coefficient in a logistic regression model depends on being able to place a meaning on the difference between two logits. The exponent of this difference gives the odds ratio, which is defined as the ratio of the odds that the independent variable will be present to the odds that it will not be present. Thus, the relationship between the logistic regression coefficient and the odds ratio provides the foundation for interpretation of all logistic regression results. It should be noted that odds greater than 1 in this study increase the likelihood that the accident will be fatal. Illustrations follow of the interpretation of the model developed in this study.

9.1. Impact of location on accident severity

It should be noted that since *LOC* has two levels as shown in Table 4, GLIM codes the first one zero and the other 1. Hence,

Location $(LOC(1)) = 0$ (Intersection)

Location $(LOC(2)) = 1$ (Non-intersection)

According to this coding, GLIM shows only *LOC(2)* in the logit model with the coefficient of 0.9697. To interpret the parameter estimate for *LOC* (0.9697), the logit difference should be computed as follows:

Logit (Fatal accident/Non-intersection)

$$= \beta_o + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$$

Logit (Fatal accident/Intersection)

$$= \beta_o + \beta_2 + \beta_3 + \beta_4 + \beta_5$$

$$\text{Logit difference} = \beta_o + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 - (\beta_o + \beta_2 + \beta_3 + \beta_4 + \beta_5) = \beta_1 = 0.9697$$

Hence the odds ratio (ψ) is

$$\psi = e^{\beta_1} = e^{0.9697} = 2.64$$

This value indicates that the odds of being in a fatal accident at a non-intersection location are 2.64 higher than those at an intersection.

Note that the logit difference (0.9697) equals the estimated value of the parameter of the independent variable *LOC* in the logit function (β_1). However, the logit difference between two levels of a dichotomous variable does not always give the parameter estimate of that variable. Since *LOC* has only two levels, the logit difference ends up with the parameter estimate. For a polytomous variable with more than two levels (or if an interaction or confounding effect exists), the logit difference is not necessarily equal to the parameter estimate. This is the case for the variable *CAUS*, shown next.

9.2. Impact of running red light on accident severity

β_2 (0.3558) measures the differential effect on the logit of two causes, $CAUS =$ run red light and $CAUS \neq$ run red light.

Table 6
A summary of *P*-values after dropping variables from saturated model

Variable dropped from the saturated model	Change in deviance	df (associated with change in deviance)	*P*-value[a]
Saturated mode	–	–	–
LIC	2.83[b]	1	0.093
VEH	1.04	1	0.31
NAT	3.36	1	0.067
CAUS	10.73	4	0.03
TIME	0.05	1	0.82
CTYP	1.38	3	0.71
ATYP	0.36	2	0.55
AGE	+0.003	1	0.452
LOC	+16.45	1	0.000[a]

[a] Based on χ^2 for the log-likelihood ratio test. For example: the *P*-value for LIC variable is obtained such as $P(\chi^2_2 \geq 3.095) = 0.213$.
[b] The +ve sign due to backward strategy. If the Forward strategy is chosen this sign would be −ve.

Table 7
The results of testing interactions and a quadratic effect of age

Variable dropped from the saturated model	Change in deviance	df (associated with change in deviance)	*P*-value
Full model[a]	–	–	–
$LOC \times CAUS$	−1.241	1	0.265
AGE^2	0.51	1	0.52

[a] Model with the two significant variables: *LOC* and *CAUS*.

To interpret this estimate, the logit difference is computed first; for example, for run red light (RRL) ($CAUS(2) = 1$), the logit is

$$\text{Logit (Fatal/RRL)} = \beta_o + \beta_1 + \beta_2$$

For any other cause but RRL, the logit is

$$\text{Logit (Fatal/Not RRL)} = \beta_o + \beta_1 + \beta_3 + \beta_4 + \beta_5$$

$$\text{Logit difference} = (\beta_o + \beta_1 + \beta_2) - (\beta_o + \beta_1 + \beta_3 + \beta_4 + \beta_5)$$

$$= \beta_2 - \beta_3 - \beta_4 - \beta_5$$

$$= -0.3558 - 0.2130 + 0.8971 + 0.6705$$

$$= 0.9988$$

Hence the odds ratio is

$$\psi = e^{\beta_2 - \beta_3 - \beta_4 - \beta_5} = e^{0.9988} = 2.72$$

Thus, the odds that an accident will be fatal because of running a red light are 2.72 times higher than for a non-RRL-related accident.

9.3. Impact of wrong way on accident severity

At a non-intersection location, the odds ratio of being involved in a fatal accident in a wrong-way-related accident are three times higher than in a failure-to-yield-related accident. This odds ratio is computed as shown above:

$$\text{Logit difference} = -0.8463 + 1.9564 = 1.1101$$

$$\psi = e^{1.1101} = 3.035$$

9.4. Odds to base level

The parameter estimates can also be interpreted in a different way for *CAUS* by relating interpretation of the estimate of any level to the base level (speed in our model). For example, the odds ratio of *CAUS(2)* can be obtained directly with no need for logit difference, as follows:

$$\beta_2 = -0.3558$$

$$\psi = e^{\beta_2} = e^{0.3558} = 0.70$$

This expression indicates that the odds ratio of the accident being fatal in an RRL-related accident is 0.70 times its being fatal in a speed-related accident, which indicates that RRL odds decrease by a factor of 0.70.

The odds ratio of either intersection or non-intersection-related accidents under different causes can be tabulated in matrix form for fast and easy interpretation, as shown in Tables 8–10. This tabulation helps to draw a conclusion for any combination of the variables in the model.

Fig. 6 presents values of the odds ratio in Table 8. It appears from this plot that a non-intersection location has greater influence on accident severity than an intersection location. One can note that all the odds for a non-intersection location are higher than those for an intersection regardless of cause. This finding indicates the odds of being involved in a fatal accident related to a non-intersection location are higher than those at an intersection. In other words, non-intersection-related accidents are more serious than intersection-related accidents in Riyadh.

Another interesting point that can be drawn from Fig. 6 is that wrong-way-related accidents exhibit significantly higher odds than do other causes. This finding means that an accident with this cause is more likely to be fatal when compared with accidents with other causes. On the other hand, failure-to-yield accidents have the lowest odds.

As shown above, the model can be used to estimate the odds ratio in order to assess the odds of an accident being fatal or non-fatal given a certain accident characteristic. This method can help in determining the most likely risk-taking behavior.

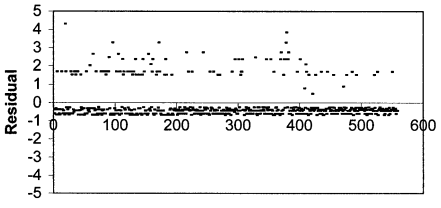

Fig. 4. Plot of Pearson residuals for graphical assessment.

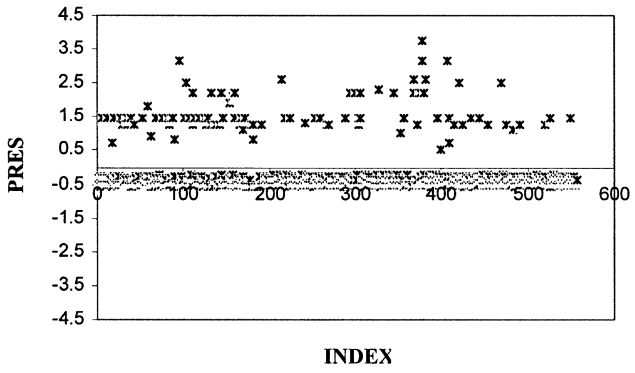

Fig. 5. High-leverage plot for graphical assessment.

10. Conclusions

Since the response variable is of a binary nature (i.e. has two categories — fatal or non-fatal), the logistic regression technique was used to develop the model in this study. The intent was to provide a demonstration of a model that can be used to assess the most important factors contributing to the severity of traffic accidents in Riyadh. On the basis of traffic police accident data, nine explanatory variables were used in the model development process.

Using the concept of deviance together with the Wald statistic, the study variables were subjected to statistical testing. Only two variables were included in the model, namely, accident location and accident cause. The observed level of significance for regression coefficients for the two variables was less than 5%, suggesting that these two variables were indeed good explanatory variables. The results presented in this paper show that the model provided a reasonable statistical fit.

Stratifying location-related data into two classes, the model revealed that the odds of a non-intersection accident being fatal are higher. This finding might lead to a greater focus on road accident sites other than intersections, which should help agencies focus their safety improvements more cost-effectively. However, it should be said that not only the relative danger as expressed by the odds ratio, but also the absolute density of accidents with regard to location should be taken into account in order to develop cost-effective strategies.

The odds presented in this paper can be used to help establish priorities for programs to reduce serious accidents. For example, since the odds of being involved in a fatal accident at a non-intersection location because of a wrong-way violation are relatively higher than those for any other violation, drivers should be warned in a specific awareness program about the possible lethality of such a violation. The same can be said of the impact of running a red light on the odds of being involved in a fatal accident.

Presentation of odds in a matrix format, as described in this study, provides a simple method for interpretation. The columns and rows of the matrix correlate the factors in the logistic model, and each cell shows the impact of a certain factor on the odds with respect to another factor (a corresponding factor).

It is important to note that the odds described in this paper were computed with no consideration for traffic exposure or the data that are not available or difficult to obtain in Riyadh. However, the findings of this study can be considered as guidance for a future study when such data become available.

Table 8
Odds of being in a fatal accident at intersection to that in a non-intersection accident[a]

Non-intersection accident	Intersection accident			
	Speed	RRL	WW	FTY
Speed	2.64	3.76	2.13	6.47
RRL	1.85	2.64	1.50	4.5
WW	3.26	4.66	2.64	8.0
FTY	1.08	1.54	0.87	2.64

[a] Example: The odds of being in a fatal accident at a non-intersection location due to WW is 3.26 times higher than that due to speed at an intersection-related accident.

Table 9
Odds of being in a fatal accident at non-intersection to that in a non-intersection accident[a]

Non-intersection accident	Non-intersection accident			
	Speed	RRL	WW	FTY
Speed	1	1.43	0.81	2.45
RRL	0.70	1	0.57	1.72
WW	1.24	1.77	1	3.03
FTY	0.41	0.58	0.33	1

[a] Example: The odds of being in a fatal accident at a non-intersection location due to WW is 1.24 times higher than that due to speed at same type of location.

Table 10
Odds of being in a fatal accident at non-intersection to that at an intersection accident[a]

Non-intersection Accident	Intersection accident			
	Speed	RRL	WW	FTY
Speed	0.38	0.54	0.31	0.93
RRL	0.27	0.38	0.21	0.65
WW	0.47	0.67	0.38	1.15
FTY	0.16	0.22	0.13	0.38

[a] Example: The odds of being in a fatal accident at a non-intersection location due to WW is far less than that due to speed at an intersection (0.47 which is less than 1).

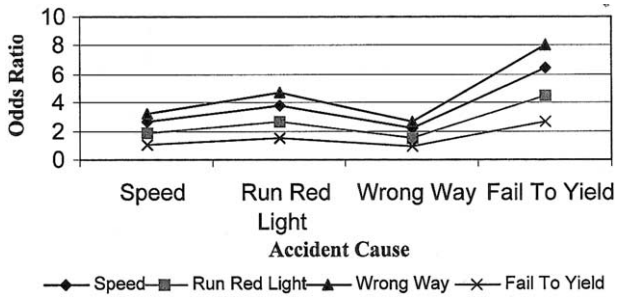

Fig. 6. Odds ratio of being involved in a fatal accident at a non-intersection location to that of an intersection relative to cause.

References

Agresti, A., 1984. Analysis of Ordinal Categorical Data. Wiley, New York.

Al-Ghamdi, A.S., 1996. Road accidents in Saudi Arabia: Comparative and analytical study. Presented at the 75th Annual Meeting of the Transportation Research Board, Washington, DC.

Feinberg, S., 1980. The Analysis of Cross-Classified Categorical Data. MIT Press, Cambridge, MA.

GLIM, 1987. Generalised Linear Interactive Modeling Manual. Release 3.77, second ed. Royal Statistical Society, UK.

Hilakivi, I., et al., 1989. A sixteen-factor personality test for predicting automobile driving accidents of young drivers. Accident Analysis and Prevention 21 (5), 413–418.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.

James, J.L., Kim, K.E., 1996. Restraint use by children involved in crashes in Hawaii, 1986–1991. In: Transportation Research Record 1560, TRB, National Research Council, Washington, DC, pp. 8–11.

Jovanis, P.P., Chang, H., 1986. Modeling the relationship of accidents to miles traveled. In: Transportation Research Record 1068, TRB, National Research Council, Washington, DC, pp. 42–51.

Kim, K., Lawrence, N., Richardson, J., Li, L., 1994. Analyzing the relationship between crash types and injury severity in motor vehicle collisions in Hawaii. In: Transportation Research Record 1467, TRB, National Research Council, Washington, DC, pp. 9–13.

Kim, K., Lawrence, N., Richardson, J., Li, L., 1995. Personal and behavioral predictors of automobile crash and injury severity. Accident Analysis and Prevention 27 (4), 469–481.

Kim, K., Lawrence, N., Richardson, J., Li, L., 1996. Modeling fault among bicyclists and drivers involved in collisions in Hawaii 1986–1991. In: Transportation Research Record 1538, TRB, National Research Council, Washington, DC, pp. 75–80.

Mannering, F.L., 1992. Male/female driver characteristics and accident risk: some new evidence. Presented at the 71st Annual Meeting of the Transportation Research Board, Washington, DC, 1992.

McCullagh and Nelder, 1982. Generalized Linear Models. Cambridge University Press, Cambridge, UK, 1982.

Mercier, C.R., Shelley, M.C., Rimkus, J., Mercier, J.M., 1997. Age and gender as predictors of injury severity in head-on highway vehicular collisions. In: Transportation Research Record 1581, TRB, National Research Council, Washington, DC.

Milton, J., Mannering, F., 1997. Relationship among highway geometric, traffic-related elements, and motor-vehicle accident frequencies. Presented at the 76th Annual Meeting of the Transportation Research Board, Washington, DC.

Nassar, S.A., Saccomanno, F.F., Shortreed, J.H., 1997. Integrated Risk Model (ARM) of Ontario. Presented at the 76th Annual Meeting of the Transportation Research Board, Washington, DC.

Official Statistics, 1997. Annual Statistics for the Period 1971–1997. Ministry of Interior, Riyadh, Saudi Arabia.