Using Logistic Regression For Estimating The Influence of Some Accident Factors on Severity

Abstract: This study applied logistic regression to accident-related data collected from traffic-police records in order to examine the contributing factors to accident severity. A total of 560 subjects involved in severe accidents was sampled. The accident severity (dependent variable) in this study is a dichotomous variable with two categories, Fatal or Injury. Therefore, each of the subjects sampled is classified as a fatal accident or an injury accident. Due to the binary nature of this dependent variable -accident severity- a logistic regression approach is suitable. Among nine variables obtained from police-accident reports, two independent variables found most significantly associated to accident severity; namely, location and cause of accident. This paper gives a statistical interpretation of the model-developed estimates in terms of odds ratio concept. The findings show that the logistic regression used in this research is a powerful tool in providing meaningful interpretations that can be used in future safety improvements in Riyadh.

INTRODUCTION
Accident severity is of special concern in traffic safety, as many efforts address accidents tend to be measures not only to prevent accidents but also to reduce the severity of accident. One way to do so is to identify the most probable contributing 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; age and, nationality of driver at fault, and his licensing status; and vehicle type. The reason for not examining more factors was due to substantial limitations of data obtained from accident reports. Logistic regression was used in this study to estimate the effect of the statically significant factors on severity. Logistic regression and other related-categorical-data regression have often been used to assess risk factors for various diseases. However, it has been also used in transportation studies. Following is a brief literature review for the use of this type of regression in traffic safety.

Regression methods have become an integral component of any data analysis concerned with describing 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, it is often the case that the outcome variable (response) is discrete. The conventional regression analysis is not appropriate. Among several reasons, the following two are the most significant:

1. The response variable in CRA must be continuous.
2. The response variable in CRA can take non-negative values.

These two primary assumptions are not satisfied when the response variable is categorical. 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 that “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. Nassar et al. developed an integrated Accident Risk Model (ARM) for policy decisions using risk factors affecting both accident occurrences on road sections, and injury severity of occupants involved in the accidents. Using negative binomial regression and a sequential binary logit formulation, the models they developed are practical and easy to use. (1997) 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 Veilahti et al. (1989) in predicting automobile driving accidents of young drivers. They examined the predictive values of the Cattel 16-factor personality 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 (199) developed a logistics regression model for describing 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 that use seat belts are far more likely to restrain their children; and one-and-two-year olds are less likely to be restrained.

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 most parsimonious. What distinguishes a logistic regression model from a linear regression model is that the outcome variable. In the logistic regression model, the outcome 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 values of the outcome variable, given the values of the independent variable as:

\[ E(Y \mid x) = \beta_0 + \beta_1 x \]

where \( Y \) denotes the outcome variable, \( x \) denotes a value of the independent variable, and the \( \beta_i \)'s denote the model parameters. The quantity is called the conditional mean or the expected value of \( Y \) given the value \( x \). Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable [Hosmer and Lemeshow, 1989]. The specific form of the logistic regression model is

\[ \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \]  \hspace{1cm} (1)

where, for simplifying notations, we let \( \pi(x) = E(Y \mid x) \). The transformation of \( \pi(x) \) logistic function is known as the logit transformation:

\[ g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x \]  \hspace{1cm} (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, may range from \(-\infty\) to \(+\infty\), dependent on the range of \( x \).
Hosmer and Lemeshow (1989) summarize the main features in a regression analysis when the outcome variable is dichotomous as follows:

1. The conditional mean of the regression equation must be formulated to be bounded between zero and 1 (equation (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 derive for logistic regression.

In linear regression the method used most often for estimating unknown parameters is least squares. In that method the values of parameters 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 the 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. For further details one can read Hosmer and Lemeshow (1989).

If \( Y \) is coded as zero or one (binary variable) then the expression \( \pi(x) \) given in equation (1) provides the conditional probability that \( Y \) is equal to 1 given \( x \). This will be 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 \( x_i \)'s are assumed to be independent, the product for the terms given in the above equation gives the likelihood function as follows:

\[
l(\beta) = \prod_{i=1}^{n} \zeta(x_i) \quad (3)
\]

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

\[
L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \left\{ y_i \ln[\pi(x_i)] + (1-y_i) \ln[1-\pi(x_i)] \right\} \quad (31)
\]

Maximizing the above function with respect \( \beta \) to and setting the resulting expressions equal to zero will produce the values of \( \beta \) as follows:

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

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

These expressions are called likelihood equations. An interesting consequence of equation (4) is that:

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

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 estimating the coefficients, the assessment of the significance of the variables in the model is taking place. If \( y_i \) denotes the observed value and \( \hat{y}_i \) denotes the predicted value for the \( i^{th} \) individual under the model, then the statistics used in the linear regression is:

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

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

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

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 to 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 equation (31). 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 following ratio:

\[ D = -2 \ln \left[ \frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} \right] \]  \hspace{1cm} (6)

is called the likelihood ratio. The reason for the minus two is that its log is mathematical. Using equation (31) and equation (6), the following test statistic can be obtained:
\[ D = -2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{\hat{\pi}_i}{y_i} \right) + (1 - y_i) \ln \left( \frac{1 - \hat{\pi}_i}{1 - y_i} \right) \right] \]  

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

The statistics \( D \), in equation (7) is called, at least in this study, deviance and 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., 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 the independent variable in the model, 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 \( G \), it can be expressed as:

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

It is not appropriate here to derive the mathematical expression of the statistic \( G \). Yet, it should be said that under the null hypothesis that is \( \beta_1 \) equal to zero, \( G \) will follow a chi-square distribution with 1 d.f. Another test statistic, similar to \( G \) for the purpose, used in this study is known as 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 as:

\[ W = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \]  

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

**Model Description**

The dependent variable in this research is Accident and of dichotomous type and stands for accident severity. Each accident in the sampled data was categorized as either injury or fatal. The logistic model used is:

\[ P(\text{Injury Accident}) = \pi(x) = \frac{e^{\beta x}}{1 + e^{\beta x}} \]  

And thus

\[ P(\text{Fatal Accident}) = 1 - P(\text{Injury}) = 1 - \pi(x) = 1 - \frac{e^{\beta x}}{1 + e^{\beta x}} \]
Where $g(x)$ stands for the function of the independent variables as:

$$
g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n
$$

Logistic regression determines the coefficients that make the observed outcome (Injury 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 to the 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 in the model. The GLIM software has built in routines to obtain Deviance and estimates of the model parameters.

**Data Description**

The data set used in this study consists of a sample of 560 subjects and was obtained from traffic police records at Riyadh (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, due to lack of qualifications and training as well as facilities needed to perform complex examinations. Even the medical reports are hard to obtain due to the fact that they are kept separately from the police reports. Police accident data and medical data are not kept together [Al-Ghamdi, 1997]. Consequently, it was impossible for this study to obtain details on the degree of severity of the accident. All that can be learned from the police records is that either the accident is a PDO accident, injury accident (no injury classification is available), or fatal accident. The subjects were selected at a systematic random process from all filed records for the period from August 1997 to November 1998. The data search was done manually due to the lack of computerization. Only injury and fatal accident records were considered for the purpose of this study. Since the goal of the study was to investigate the factors which might affect the consequence of the accident (i.e., either fatal or injury accident), ten variables were summarized from the data. The description and codes of these variables are listed in Table 1. Figure 1 shows the age distribution for drivers in the data set.

**Please Insert Figure 1 about here**

**Please Insert Table 1 about here**

The response variable is variable number 1, namely, Accident (ACCIDENT) which is binary (dichotomous) in nature. Two levels for ACCIDENT exists: 0 if the accident results in at least one injury but no fatality, and 1 when there is at least one fatality resulting from the accident. For the explanatory variables (independent variables), age is the only continuous variable and the others are categorical. Since some of the categorical variables have several levels and there should be an identifier (number: 1,2,3,...), 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.

One possible way of coding 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, there should be three design variables. When the respondent is “With vehicle(s)” the three design variables, $D_1$, $D_2$, and $D_3$ would be all set to equal zero; when the respondent is “Fixed-object”, $D_1$ would be set equal to 1 while $D_2$ and $D_3$ would still equal 0; and so forth for the other respondents. This coding
scheme was done for the rest of the categorical variables. It should be said that GLIM has the capability to do this coding scheme automatically once levels of variables are identified by the end-user. Other software packages would have different strategies of coding design variables. It is important to understand the coding strategy in the software package used in order to conduct testing hypothesis on the variables as well as for interpretation of their estimates.

Reducing Design variables
As can be seen from Table 1 that some of the categorical variables have several levels so that several design variables are needed for each. Generally speaking, it is more convenient to have as minimum as possible of design variables in order to simplify the model interpretation. In other words, the more design variables the model includes, the more difficult the interpretation becomes. Thus, the attempt was made at early stages of this study to reduce the design variables. However, care is needed to do so to guarantee that the model will not loose significant information.

Looking at the proportions of level for the study variables as shown in Figure 2, one can see that some levels can be neglected due to small proportions. However, eye investigation is not enough to decide which levels can be neglected or at least can be 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 can be reduced or not. The following typical test is used

\[ H_0 \quad p_i = 0 \]
\[ H_a \quad p_i \neq 0 \]

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

For example the design variables for ATYP was reduced from 3 design variable (4 levels) to two design variables (three levels) after showing that the proportions of "fixed-object" and "overturn" accidents were not statistically significant at 5% level of significant using the above 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 the testing.

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

Developing the Logistic Model
The backward selection process of logistic regression was followed in this study. First, all of the variables with no interactions (referred to here as a saturated model (Figure 3)) were tested
based on Wald (W) statistic and Deviance defined in equation (7) and equation (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 column of 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) statistics to assess the significance of the coefficients. For example, the variable CAUS has 5 levels (Figure 2) but only two of the levels (CAUS(5)) was found statistically significant at 0.05 level. In this case, the decision to include this variable should be made using the likelihood ratio test. That is, the change in deviance for the model, with the variable and without it, should be assessed.

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-value=0.0925) as presented in Table 6. This indicates that LIC is not adding useful information to the variability in the response variable and should be removed. Similarly, the variables VEH, TIME, CTYPE, ATYPE, AGE, and NAT do not show any major changes in the deviation, 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 three variables (LOC and CAUS) as being related significantly to accident severity. These two variables were then subjected for further analysis as will be seen shortly. Before that, one might believe that accident type (ATYP) and collision type (CTYP) should have a significant effect on the accident severity, yet, not true in this study since they failed to meet the desirable level significance (0.05). However, one might argue that these two variables could be implied in the two significant variables in the model, namely, LOC and CAUS. For example, knowing that severe accident occurs at an intersection, right angle would be the most likely collision type caused most often by running red light (note that rig-angle type and red-light cause are significant proportion in Table 3). Right-angle collision due to running red light is a very 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) would imply a multiple vehicle, fixed-object or pedestrian accident (ATYP).

Interaction and Confounding Effects

The two variables (i.e., LOC and CAUS) that were found to be statistically significant in the current study 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 interactions was found statistically insignificant (p-value is 0.265), as presented in Table 7, and hence, confounding effect does not exist.
Age Effect
Understanding and quantifying the relationship between driver characteristics, in particular age and accident risk, has long been a high priority of accident-related research. Additionally, other studies (e.g., Hilakive et al., 1989; Mercier et al., 1997) have shown that young drivers, as well as older drivers, are more at risk for being involved in severe accidents. Numerous research studies have made attempts to examine the complex relationship between driver characteristics and accident risk. Drivers’ risk-taking behavior is often defined in terms of several variables among which one is age. Mannering (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 with the response variable [Mercier et al., 1997; Hosmer and Lemeshow, 1989]. They suggest a quadratic expression.

The problem with the age variable in this study appears from the unexpected positive effect shown in the parameter estimate as shown in Table 5. It was expected that the older the driver the less 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 more closely investigate the age factor, even though it has been shown from the analysis in this study that age is not statistically significant. Hence, the age variable was subjected to more examination. The model has shown so far that age, in a linear relation form 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 is not statistically significant either (p-value is 0.52, as in Table 7).

The Logit Model
According to the previous analysis, the logit model with the significant variables is:

\[
\hat{g}(x) = -2.029 + 0.9697 \text{LOC}(2) - 0.3558 \text{CAUS}(2) + 0.2130 \text{CAUS}(3) - 0.8971 \text{CAUS}(4) - 0.6705 \text{CAUS}(5)
\]

Hence the logistic regression model developed in this study is:

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

Testing for the Significance of the Model
Once the model has been fit, the process of assessment of the model begins. Several tests, including Pearson chi-square and deviance, Wald Statistic, and the Hosmer-Lemeshow Tests, can be used to determine how effective the model is in describing the outcome variable. This is referred as its goodness-of-fit. These tests end up with a chi-square criterion to make the decision on the model fit. A very good reference 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 Wald statistic, as discussed early.

Graphical assessment of fit to the logistic model developed in the study also shows that the model appears to fit reasonably the data as shown in Figure 4 and Figure 5. Figure 4 shows the plot of Pearson residuals in which no trend can be detected. The other plot in
Figure 5 shows Hi-Leverage points (outliers) in which very small points seem to be outliers (less than 4% of data set: compare PRES with 1.96 (z-value at 5% level of significance). That is, 95% of the points in this plot lay within –0.5 and 1.9.

MODEL INTERPRETATION

The 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., link function [McCullagh and Nelder, 1983]), and appropriately defining the unit change for the independent variable. In the logistic regression model, the link function is the logit transformation (eqn. (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 meaning on the difference between two logits. The exponent of this difference (the difference between the two logits) gives the Odds Ratio which is defined as the ratio of the odds for the independent variable being present to the odds of not being present. Thus, the relationship between the logistic regression coefficient and the odds ratio provide the foundation for interpretation of all logistic regression results. It should be said that odds greater than one in this study increase the likelihood of accident to be fatal. Following subsections give illustrations for interpretation for the model developed in this study.

Impact of Location on Accident Severity

The estimate for LOC is 0.9697 can be transferred to the odds ratio as:

It should be said that since LOC has two level as shown in Table 4, GLIM gives the first one the code zero and 1 for the other level. 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. To interpret the parameter estimate (0.9697) for LOC, the logit difference should be computed as follows:

\[
\text{Logit (Fatal Accident/Non-intersection)} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 \\
\text{Logit (Fatal Accident/Intersection)} = \beta_0 + \beta_2 + \beta_3 + \beta_4 + \beta_5
\]

\[
\text{Logit difference} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 - (\beta_0 + \beta_2 + \beta_3 + \beta_4 + \beta_5) \\
= \beta_1 \\
= 0.9697
\]

Hence the odds ratio (\( \psi \)) is

\[
\psi = e^{\beta_1} = e^{0.9697} = 2.64
\]

This indicates that the odds being in a fatal accident at a non-intersection is 2.64 higher than at an intersection.
Note that the logit difference (0.9697) equals the estimate value of the parameter of the independent variable LOC in the logit function ($\beta_1$). However, it is not always the case that the logit difference between two levels of a dichotomous variable gives the parameter estimate of that variable. Since LOC has only two levels, the logit difference ends up with the parameter estimate. For a dichotomous variable with more than two levels (or if interaction or confounding effect exists), the logit difference is not necessarily equal to the parameter estimate. This case can be illustrated in the next illustration for the variable CAUS:

### Impact of Running Red Light on Accident Severity

$\beta_2$ (0.3558) measures the differential effect on the logit of two causes, CAUS = Run red light and CAUS $\neq$ Run red light.

To interpret this estimate we should first compute the logit difference such as

For Running Red Light (CAUS (2) = 1), the logit is

$$\text{Logit (Fatal/Running Red Light)} = \beta_0 + \beta_1 + \beta_2$$

For any other cause but not Running red light, the logit is

$$\text{Logit (Fatal/Not Running Red Light)} = \beta_0 + \beta_1 + \beta_3 + \beta_4 + \beta_5$$

The logit difference = $(\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_1 + \beta_3 + \beta_4 + \beta_5)$

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

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

$$= 0.9988$$

Hence the odds ratio is

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

Thus, odds of a Accident being fatal due to running red light is 2.72 times higher than that in a non-running red light-related accident.

### Impact of Wrong Way on Accident Severity

In a non-intersection location, the odds ratio of being involved in a fatal accident in a wrong way-related accident is 3 times higher than that in a failure-to-yield related accident. This odds ratio is computed as shown above:

Logit difference = -0.8463 + 1.9564 = 1.1101

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

### Odds to the Base Level

Also the parameter estimates can be interpreted in a different way for CAUS by referring interpretation of the estimate of any level to the base level (Speed in our model). For example, we can obtain the odds ratio of CAUS(2) directly with no need for logit difference, as follows:

$$\beta_2 = -0.3558$$

$$\psi = e^{\beta_2} = e^{-0.3558} = 0.70$$
This indicates that the odds ratio of the Accident being fatal in running red light-related accident is 0.70 times that being fatal in a speed-related accident, which indicate that RRL odds decrease by a factor 0.70.

The odds ratio of either intersection or non-intersection related accidents under different causes can be tabulated in matrices form for fast and easy interpretation, as shown in Tables 8, 9, and 10. This tabulation helps to draw a conclusion of any combination of the variables in the model.

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

Another interesting point can be drawn from figure 6 that wrong-way related accidents exhibit significantly higher odds than other causes. This means that the accident of this cause has more likely to be fatal when compared to other causes. On the other hand, failed-to-yield accidents have the lowest odds.

Please Insert Table 8 about here
Please Insert Table 9 about here
Please Insert Table 10 about here
Please Insert Figure 6 about here

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

CONCLUSIONS
Since the response variable is of binary nature (i.e., has two categories: fatal or injury), 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 contributing factors to the severity of traffic accidents in Riyadh. Based on traffic police data, nine variables were used as explanatory variables in the development process.

Using the concept of Deviance together with 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 were 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 by into two classes, the model revealed that odds of a non-intersection accident to be fatal are higher. This might lead to say that this more enforcement focus should be given to road site far from intersections.

Analysis of odds in this study also shows that accidents due to wrong-way driving are more likely to be fatal.

The findings of the study should help law enforcement agencies focus not only their enforcement efforts but also their awareness campaigns to establish priorities in order to
reduce accident severity. For example, drivers should be aware of the fatal fate of being involved in a wrong-way-driving accident.

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

REFERENCES


Jovanis, P. and Chang, H. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1086, 42-51.


