Available online at www.CivileJournal.org

Civil Engineering Journal

Vol. 1, No. 1, November, 2015

# An Examination of Crash Severity Differences Between Male and Female Drivers, Using Logistic Regression Model

Alireza Pakgohar[a], Mojtaba Kazemi[b*]

[a] *PhD Student, Department of Statistics, International Campus of Ferdowsi University of Mashhad, Iran*.

[b]*MSc. Highway and Transportation engineering, Department of civil engineering, Roudsar and Amlash branch, Islamic Azad University, Roudsar, Iran*

*Received 18 October 2015; Accepted 30 November 2015*

**Abstract**

One person in every 2539 people gets killed and one in every 253 suffers injuries due to driving crashes each year in Iran. Such that driving incidents are second rank factor of death and the first rank reason for lost lifetimes in this country. 60% of total incidents which lead to deaths or injuries are actually driving incidents in Iran. That is while the same ratio is only 25% worldwide average. In this article, we report a probabilistic relationship between vehicle drivers' gender and severity of the accidents. The model accuracy rate is more than 91%. Coefficient values show that if an crash happens and all other variables are under control, the probability of suffering injuries for a man is 1.597 times more than for a woman (1.40 – 1.79, 99% CI) in comparison with the case that the person does not get injured at all. Similarly, the probability of death for a man is 1.462 times higher than for a woman (1.13-1.79, 90% CI) again in comparison with case of no injury at all.

*Keywords:* Gender, Road Crashes, Crash Severity, Logistic Regression.

## 1. Introduction

Taking into account the 6,342,000 population of the world in 2004, one person in every 5,285 dies and one in every 127 people suffers injuries due to driving incidents each year. These figures are 2,539 and 253 respectively (Pakgohar, 2012). Total yearly direct and indirect costs imposed by driving crashes amount to 180,000 billion Rials. This estimate cost amount is equal to 6.23% of GDP of the country in 2007. Since the GDP growth rate was 6.7% that year it can be concluded that driving crashes cost swallows almost all the growth of GDP in Iran. A statistical report from Health Ministry shows that driving crash is the second rank cause of death and first rank cause of lost lifetime in Iran (average world statistics show this factor in rank nine). 60% of total incidents which cause death or injury in this country are actually driving incidents, while worldwide average is only 25% (Pourmoalem and Ghorbani, 2011). Humans are different in terms of physical, psychological, social, and recognition abilities. This is true in driving as well; Such that people with higher sensory skills, lower reaction time, and higher precision are more successful in driving. Sensation seeking is a personal characteristic which influences peoples driving behavior. The person in this case tends to experience new things and risks for them. Males and females are different in this sense (Soori, 2005). Significant differences have been observed in other traits like intelligent cognition and sensation (Esmaily, 2010).

This research addresses the effect of person's gender on crash severity. Binary logistic regression method is used in this research to tackle this job. The level of probability for occurrence of some situation can be determined using this specific method.

[*] Corresponding author: mojtaba.kazemi88@yahoo.com

## 2. Research background

The primary goal of the research was to examine how human factors influence crash severity prediction and categorization in Iran. Data regarding crashes happened in 2007 were used for this task. Results obtained using tree regression and logistic regression methods suggested that having driving license, using seat belt, age, and gender all influence the severity of road crashess as human factor indices (Pakgohar et al., 2011).

Waylen and McKenna (2000) have shown that crash involvement patterns are different for two genders. Men are more probable to get involved in crashes on the bends, low light or overturn situations, while women on the other hand are more prone to get involved in crashes at the intersections and junctions than men (Waylen 2000)

Men perform better on assessing time intervals. Although women recognize the movement faster than men; but they estimate the distance shorter. These differences are reasons why female drivers keep longer distances and break faster and bore severely at emergencies. As the studies have shown, women react faster than men to dangerous situations, but sometimes act to control the vehicle, for example turn the steering wheel and apply breaks with more delay in comparison to men, because of physical conditions (Leen, 2004). Landaur and colleagues (1980) believe that on average, women react faster than men on a time task. Reaction time average was 0.485 sec. for women and 0.534 for men. Parker and Lajunen, (2001) study has shown that aggressive driving is far more seen in men than women (Esmaily, 2012).

## 3. Methodology

Usually it takes to use logistic regression when multi-value data are to be processed as dependent variables. Especially, this method is more prominent in organic assay and audit analysis. On the other hand, since prediction by logistic regression is in fact some type of classification, we can set audit analysis in this framework as well. Logistic regression is one of the most applicable generalized linear models used to analyze relations between one or more descriptive variable/s and a scalar response variable. Therefore, Logistic regression opens wider fields of statistical analysis before people's eyes (Mojtaba Kazemi, 2011).

Our methodology is categorized as "Descriptive Research" in the realm of social studies and "descriptive-analytic" in terms of viewpoint and problem addressing. The statistical analyses used in this paper include statistical descriptive measures such as average, percents, etc… and Logistic regression (LR) model.

We use LR as the primary model to recognize patterns of crash severity applied to the driver based on driver gender. Many papers have used Logit model (for example, Pakgohar and Khalili (2010), Pakgohar et al., (2011), and Esmaeili et al., 2012). There are a number of reasons to use this method. First, the Logit model has been widely used and well developed. Second, it is relatively easy to understand and is integrated readily in most software packages. Our last reason is that the Logit model is well known as an accurate and reliable tool for predictions.

Logistic Regression mode is a nonlinear transformation of linear regression model (LN transformation). The logistic distribution is an S shaped distribution function similar to standard normal distribution. Like multivariate regression, researchers are interested to find a suitable arrangement for predictor variables which helps with interpreting binary results.

With logistic regression, the probability of occurrence for a certain event is directly estimated. In case only one predictor exists, logistic regression can be formulated as:

$$\text{Occurrence probability} = f(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (1)$$

In which $\beta_0$ and $\beta_1$ are coefficients that would be estimated using data (original samples), and *x* is the predictor. The formulation with more than one predictor variable is as follows:

$$\text{Occurrence probability} = \frac{e^{z}}{1+e^{z}} \text{ or } \frac{1}{1+e^{-z}} \qquad (2)$$

In which$= \beta_0 z_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_k z_k = \beta_0 + \sum_{i=1}^{n} \beta_{i_\sim} x_i$

Clearly the probability for event not to occur is $\frac{e^{z}}{1-e^{z}}$. These relations are called multi-variable logistic functions. A linear pattern at Logit scale would be fitted using Logit transformation introduced above:

$$\log\ it\ (\pi_i) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (3)$$

So Logit modeling in respect to $\pi_{\mathrm{i}} = {}^{\mathrm{y_i}}/_{\mathrm{n_i}}$ is regarded as a linear function of predictors. Equation (4) can be obtained from this relation (Mojtaba Kazemi, 2011).

$$E\left(\frac{y_i}{n_i}\right) = \pi_i = \left[\frac{\exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}{1+\exp(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}\right] \qquad (4)$$

The accident severity applied to the driver is represented by a random binary variable Y in this study which would be given as:

$$Y = \begin{cases} 1 & \mathit{if\ driver\ is\ injured} \\ 2 & \mathit{if\ driver\ is\ dead} \\ 0 & \mathit{if\ driver\ is\ alive} \end{cases}$$

In which:

$$\mathrm{P(Y = 1|x) = h(\beta'X)} \qquad (5)$$

Although Logit model is called non-parametric, but function h postulates are totally parametric in statistical deduction. Especially h is a logical accumulated distribution function and is formulated as:

$$h(\beta'X) = \frac{\exp(\beta'X)}{1+\exp(\beta'X)} \qquad (6)$$

### 3.1. Model criteria

- **Model test**

Likelihood ratio test for the main model - which is called chi square test as well - is used for comparing researcher's model against a trivial model as a basis with a constant value. Chi square likelihood ratio test will be given by deducting deviation (-2ll) of final (complete) model from deviation of "sheer intersection" model. Number of degrees of freedom would be equal to number of terms minus 1 for this test. (Munizaga and Alvarez-Daziano, 2005)

- **Gauges of data fit information model**

Biesian information criterion (BIC) and Akaike coefficient or AIC are general information theory statistics and are used when we want to compare alternative models. Lower value for them indicates better fit for the models (Munizaga and Alvarez-Daziano, 2005).

### 3.2. Model efficiency measures

1. **Classification Accuracy:** This measure shows the ratio of correct predictions over positive and negative input. This measure is largely dependent on dataset distribution, and therefore can easily lead to wrong results regarding system efficiency.

2. **Classification Sensitivity:** This measure evaluates the ratio of true positives, i.e. gives the extent of system ability to predict correct values out of total input items.

3. **False positive ratio:** In binary regression, the number of wrong predictions in which the dependent variable is predicted to have value 1, but it really has the value 0. This ratio is stated as a percent of total observations. In multivariate regression, the number of wrong predictions for which the predicted value of the variable is higher than actual observed value. This is stated as a percent of total items on or above diagonal.

4. **False negative ratio:** In binary regression, the number of wrong predictions in which the value of 0 is predicted for dependent variable, but the actual observed value is 1. This is stated as a percent of all observations with value 1. In multivariate logistic regression, the number of wrong predictions in which the predicted value for dependent variable is less than observed value. This measure is stated as a percent of total number of items on or below diagonal (Munizaga and Alvarez-Daziano, 2005).

## 4. Results

This cross sectional article is based on database COM 114* information sources. Data for accidents on 2006-2007 are used which were recorded by "Rahvar" Police officers. Size of the statistical sample was 376,170 situation sketches made from road crashes in Iran. Around 10% of data were omitted during cleaning process, and regression model was fitted with the rest of data. Research variables are driver gender (Table 1) and his/her health condition after crash (Table 2). Descriptive analysis of research database showed that around 91 percent of people involved in a crash

* Iranian Guidance and Driving Police Road Accident Database

were not injured; 8 percent of crashes lead to injuries and 1 percent was fatal. We can say, based on the obtained information the number of crashes causing injury is 8 times the number of fatal ones (Table 2).

**Table 1. Descriptive Statistics, drivers' gender**

| | Label | Frequency | Percent | Valid Percent |
|---|---|---|---|---|
| Male | SEX_ID=1 | 335031 | 89% | 99% |
| Female | SEX_ID=2 | 3663 | 1% | 1% |
| Unknown | Missing Data | 37476 | 10% | - |
| Total | | 376170 | 100% | 100% |

**Table 2. Descriptive Statistics, crashes severity**

| | Label | Frequency | Percent | Valid Percent |
|---|---|---|---|---|
| No Injury | IMPACT_TYP = 1 | 307995 | 82% | 91% |
| Injury | IMPACT_TYP = 2 | 26887 | 7% | 8% |
| Fatal | IMPACT_TYP = 3 | 3256 | 1% | 1% |
| Unknown | Missing Data | 38032 | 10% | - |
| Total | | 376170 | 100% | 100% |

A Multivariate Logistic Regression analysis was conducted in which the severity of injuries taken by the driver was selected as dependent variable and his/her gender as independent variable (predictor). In total, more than 338 thousand people were included in analysis and Full model was significantly stable ($P - value < 0.0001,\ df = 2$ and $chi\ square = 48.268$). AIC and BIC coefficients are given in Table (3). Since the measures like R2 show the strength of effect along with fitting quality and cause confusion, we used correctness index in ranking table to determine accuracy of the model. Model accuracy was 91.1% on this basis; model sensitivity index was 100%, False Positive rate was zero and False Negative rate was 8.9% (Table 4).

**Table 3. Model fitting information**

| Model | Model Fitting Criteria | | | Likelihood Ratio Tests | | |
|---|---|---|---|---|---|---|
| | AIC | BIC | -2 Log Likelihood | Chi-Square | df | Sig. |
| Intercept only | 86.207 | 107.669 | 82.207 | | | |
| Final | 41.939 | 84.863 | 33.939 | 48.268 | 2 | 0.000 |

**Table 4. Classification**

| Observed | Predicted | | | |
|---|---|---|---|---|
| | 1.00 | 2.00 | 3.00 | Percent Correct |
| 1.00 | 307994 | 0 | 0 | 100.0% |
| 2.00 | 26887 | 0 | 0 | 0.0% |
| 3.00 | 3256 | 0 | 0 | 0.0% |
| Overall Percentage | 100.0 % | .0% | .0% | 91.1% |

WALD coefficient and statistic and related degrees of freedom along with probabilities for each level of the predictor variable are presented in Table (5). This shows that gender can predict ultimate state of driver injuries due to accident with 99% reliability. The confidence of prediction is about 90% in fatal cases. The coefficient values show that likelihood for a person to get injured (compared to him/her getting no injuries at all) during crash is 1.597 times higher (1.40-1.79, 99% CI) for a man compared to a woman if all other variables are controlled for. The likelihood of a person to die during crash (compared to him/her getting no injuries at all) is 1.462 times higher for a man (1.13-1.79, 90% CI) than a woman.

**Table 5. Parameter Estimates**

| IMPACT_TYP[3] | B | Std. Error | Wald | df | Sig. | Exp (B) |
|---|---|---|---|---|---|---|
| 2.00 Intercept | -2.903 | 0.075 | 1509.546 | 1 | | |
| [SEX_ID=1.00] | 0.468 | 0.075 | 39.023 | 1 | 0.000 | 1.597 |
| [SEX_ID=2.00] | 0[b] | | | 0 | | |
| 3.00 Intercept | -4.926 | 0.201 | 602.146 | 1 | | |
| [SEX_ID=1.00] | 0.379 | 0.202 | 3.546 | 1 | 0.000 | 1.462 |
| [SEX_ID=2.00] | 0[b] | | | 0 | | |

a. The reference category is: 1.00
b. This parameter is set to zero because it is redundant.

## 4. Conclusions

A logistic regression model in this study was able to reach a very high accuracy (91%) in predicting effective cause on road crash severity factor. Based on fitted model, chance of a male driver to get injured during a crash is around 8%; and this likelihood for a female driver is 0.05. Similarly, likelihood of death for a male driver who gets involved in an accident is 1% and that likelihood for a female driver is 0.7%.

It seems that other field factors such as speed factor effect during crash are statistically significant. On the other hand, speed factor has been considered in an average form for both gender groups, so we cannot consider it as a definitive factor. Other factor like vehicle room atmosphere could be considered which provides safer environment for females when occupant bumps into interior of the vehicle. Smaller and shorter body and less weight (lower body mass) will keep female drivers safer. We suggest this study to be done on male and female drivers with identical body statistics as well. Unfortunately available statistical data did not provide information regarding vehicle speed at the time of crash, driver's weight, age and height, crash type (vehicle – vehicle) and vehicle type. It seems that tree regression can classify people into homogenous groups when high volumes of data are used.

## 5. References

[1]. Alireza Pakgohar, Mojtaba Kazemi,. Determine the effect of the components of the error in traffic accidents, Scientific-reseasrch Quarterly of Motaleat-e Pajooheshi Rahvar. Vol 1, No 3 (2013):115-142.

[2]. Pourmoalem, N., Ghorbani, M. "Roads safety Portrait", Road Ministry pub. Press, national road safety commission (2011).

[3]. Soori, A., "Tehran Traffic psychology", Police Science University (2005).

[4]. Esmaili, A., Male and female driving differences in Tehran, Police Science University (2010).

[5]. Pakgohar, Alireza, Reza Sigari Tabrizi, Mohadeseh Khalili, and Alireza Esmaeili. "The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach." Procedia Computer Science 3 (2011): 764-769.

[6]. Waylen, A. and McKenna, F. Cradle Attitudes-Grave Consequences. The development of gender differences in risky attitudes and behavior in road use. UK, Basingstoke: AA Foundation for Road Safety (2000).

[7]. Leenbe, Frik, "Accident scene recreation", translated by Abbas Razanif, Soroush Publication press, Police Science University (2004).

[8]. Landauer, Ali A., Simon Armstrong, and Joanne Digwood. "Sex difference in choice reaction time." British Journal of Psychology 71, no. 4 (1980): 551-555.

[9]. Alireza Pakgohar, Mojtaba Kazemi, "Evaluation of driver behaviour towards traffic signs" Master degree thesis. Azad Islamic University of Zanjan, Iran (2011).

[10]. Khalili, Mohadeseh, and Alireza Pakgohar. "Logistic Regression Approach in Road Defects Impact on Accident Severity." Journal of Emerging Technologies in Web Intelligence 5, no. 2 (2013): 132-135.

[11]. Munizaga, Marcela, and Ricardo Alvarez-Daziano. "Testing mixed logit and probit models by simulation." Transportation Research Record: Journal of the Transportation Research Board 1921 (2005): 53-62.