Predictive Probability Models of Road Traffic Human Deaths with Demographic Factors in Ghana

Christian Akrong Hesse
  University of Professional Studies Accra, Ghana

Dominic Buer Boyetey (dbboyetey@uesd.edu.gh)
  University of Environment and Sustainable Development, Somanya, Ghana

Albert Ayi Ashiagbor
  University of Professional Studies Accra, Ghana

Research Article

Keywords: Age groups, gender, linear model, logistic regression, probit regression, road traffic fatalities

Posted Date: June 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1726381/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Predictive Probability Models of Road Traffic Human Deaths with Demographic Factors in Ghana

Christian Akrong Hesse ¹, Dominic Buer Boyetey ² and Albert Ayi Ashiagbor ¹

¹Department of Banking and Finance, University of Professional Studies Accra, Ghana
²Department of Sustainable Energy and Resources, University of Environment and Sustainable Development, Somanya, Ghana

Corresponding Author:
Dominic Buer Boyetey (PhD)
Email: dbboyetey@uesd.edu.gh
ORCID #: 0000-0002-4411-3256

Abstract
Road traffic carnages are global concerns and seemingly on the rise in Ghana. Several risk factors have been studied as associated with road traffic fatalities. However, inadequate road traffic fatality (RTF) data and inconsistent probability outcomes for RTF remain major challenges. The objective of this study is to illustrate and estimate probability models that can predict road traffic fatalities. We relied on 66,159 recorded casualties who were involved in road traffic accidents in Ghana from 2015 to 2019. Three generalized linear models, namely logistic regression, probit regression and linear probability model were used for the analysis. We found that gender and age groups have significant effects in predicting the probability of road traffic fatality for all three models. Through a likelihood ratio test, however, it was determined that the logit regression model produced consistent probabilities of traffic fatalities which are very close to the actual probability values across the age groups and gender, compared to the other two models. Thus, we recommend intensified campaign for the use of seat belts in vehicles, targeted at the aged and male users of road transport to reduce the possibility of death in any RTA.

Keywords:
Age groups, gender, linear model, logistic regression, probit regression, road traffic fatalities

Introduction
Reducing risks of road traffic fatalities (RTF) remains the ultimate objective in many road safety regulations and studies. Deaths resulting from road traffic crashes have become an existential threat, as available statistics indicate that the world loses close to 1.35 million
people through road accidents each year, of which majority are young people between the ages of 5 to 29 years (Ullah, Farooq & Shah, 2021). The repercussions are severe given the immense losses they bring to the victims’ families and their communities. Further, it is estimated that several countries make economic losses of about three to five percent of their gross domestic product (GDP) due to road traffic crashes (World Health Organization [WHO], 2018). The situation persists as the world missed the Sustainable Development Goals (SDG) target 3.6, to half road traffic deaths by 2020 (United Nations [UN] GA/10920, 2020).

This notwithstanding, the UN General Assembly in its resolution 74/299 on improving global road safety, reemphasized the target of halving the global number of deaths and injuries from road traffic crashes by 2030 (UN, 2020). To this end, examining the risks of road traffic fatalities by victim’s characteristics is crucial to comprehensively understand the impact of various risk factors contributing to the fatalities so that appropriate safety interventions can be identified and implemented to reduce the number of deaths from these crashes. Road traffic accident fatalities, according to WHO (1979), include only deaths which occur within 30 days following a road accident.

Statistics available from WHO (2018) suggest significant differences in the rate of road traffic deaths per 100,000 people across different regions of the world. Among these regions (Africa, America, Eastern Mediterranean, Europe, South-East Asia and Western Pacific), Africa was noted to have the highest rate of road traffic deaths (26.6/100,000 people) compared to Europe (9.3/100,000 people). Further evidence from OECD (2020) mentioned South Africa to have had a road traffic mortality rate of 22.4/100,000 people in 2018. These figures present an unpleasant trend in terms of the progress made in fighting road traffic carnages in Africa. Particularly in Sub-Saharan Africa, Aga et al. (2021) explained that the situation of road traffic accidents is severe, and that the region has the highest road traffic death rate, with significant number of properties damaged through road traffic accidents.

The evidence in Ghana is frightening as Blankson and Lartey (2020) confirmed that deaths resulting from road traffic accidents constitute 62 percent of all emergency cases reported at designated referral hospitals for accident victims in Ghana. This was corroborated by available statistics from WHO (2020), which indicates an average of 8 persons in Ghana out of every 100,000 population die from RTA annually over the past decade (Blankson & Lartey, 2020). Many of these RTA related deaths, according to Konlan et al. (2020), are caused by road traffic behavior of motorcyclists. Predictor variables such as age, alcohol influence, excessive speeding, bad roads, overloading and disregard for road regulations have
been studied to be significant risk factors of road fatalities in Ghana (Agyemang, Abledu, & Semevoh, 2013; Siaw, Duodu, & Sarkodie, 2013; Nyamuame, Aglina, Akple, Philip & Klomegah, 2015; Asare & Mensah, 2020). However, these studies are devoid of vigorous probability models that attempts to predict road traffic human deaths in Ghana. For instance, Asare and Mensah (2020) only applied the ordinal regression model to identify factors that contribute to accident severity in Ghana.

**Literature Review**

This section discusses contributory risk factors to RTF and methodological approaches used in estimating these risk factors. It further presents a brief empirical review of RTF studies across the globe. Extensive research has already provided many insights into risk factors that influence road traffic crashes. The literature discussions have focused on six areas; demographic factors, human factors, road factors, vehicle factors, circumstantial factors and environmental factors. Demographic characteristics such as gender, age, education, employment sector and income earned by drivers dominate recent studies (Mehdizadeh, Shariat-Mohaymany & Nordfjaern, 2018; Machado-Leon et al., 2016), but road traffic crashes resulting in deaths are not exclusive to only drivers of the vehicle. Other studies (Regev et al., 2018; Mishra et al., 2010) show that RTF cuts across different spectrum of road users with different demographic backgrounds.

Regev et al. (2018) and Melchor et al. (2015) noted that men are more likely to suffer death in a road accident because of their higher frequency in engaging in the transport and distribution business compared to women. In relation to age, Vaa (2003) observed that the biological and psychological system of a person deteriorate faster as they age. This makes the aged (60 years and above) more likely to suffer death in a road accident. Vaa’s (2003) observation is in contrast with Hesse and Ofosu’s (2014) study where between the periods of 2001 to 2010, they found that persons between the ages of 26 to 35 were the highest casualties in road traffic fatalities in Ghana.

On the part of human factors, Wu and Xu (2017), Rolison et al. (2018) and Abele et al. (2018) highlighted that driver behaviors such as speeding, drunk-driving, fatigue, safety measures adopted and risk-taking behaviors are the most influential causes of traffic-related causalities. Febres et al. (2019) blamed young drivers for risky driving behaviors. Mazankova (2017) maintained that these human behaviors are the main cause of about 70 percent of the road traffic fatalities. Zhang et al. (2009) and Yau, Zhang and Li (2016) found road related factors such as increased motorization, lane changing and overtaking cars as having negative effects on traffic safety. Lack of using appropriate safety accessories in vehicles, according to
Febres et al. (2019), contribute a higher probability of human deaths on motorways. The environmental factors including road types, nighttime travel and weather conditions were examined by Altwajri et al. (2011) to have amplified the risk of exposure to fatal road injuries causing deaths.

The literature presents a plateau of statistical models used in predicting human road traffic fatalities and injuries given some risk factors. In Farooq and Moslem’s (2020) studies, analytic network process was used to conclude that driving without alcohol and obeying speed limits were significant factors compared to other factors causing road traffic injuries in Hungary. This conclusion is problematic as the study is limited in its sampling. Twenty drivers were used without regards to any probabilistic approaches; besides, analytic network process is only a decision analysis tool. The analytic network process functions like the artificial neural networks, as applied in similar previous studies such Delen et al. (2006) and Chimba and Sando (2009). A related study by Febres et al. (2019) used Bayesian network to conclude that lack of using appropriate safety accessories, high speed violations, distractions as well as errors have higher probability of predicting fatal injuries for drivers in Spain. Febres et al.’s (2019) study used secondary data where 66,253 drivers were selected using systematic sampling compared to Farooq and Moslem’s (2020) study that gathered data through a questionnaire survey.

The use of Bayesian statistics in predicting risk of RTF is increasing. Varied approaches such as Bayesian ordered probit and Bayesian hierarchical binomial logit are common in the literature (Xie et al., 2009; Huang et al., 2008). For instance, Hesse et al. (2014) used Bayesian analysis to confirm that population and numbers of registered vehicles were the predominant factors influencing road traffic fatalities in Ghana. The weakness of the Bayesian approach is however exposed mainly in its subjective choice of priors.

Another statistical concept common in RTF literature is the logistic regression analysis. It is presented in three forms, namely the binomial (binary), multinomial and ordinal logistic regression. Santos, Picado-Santos and Trindade (2017) applied the concept of the binary logistic regression to identify the factors influencing work zone road crashes. Other studies that employed the binary logistic regression model in RTF analyses include Potoglou et al. (2018), Zeng et al. (2019) and Ebolia, Forcinitia and Mazzullaa (2020). The major weakness of the binary logistic regression is that it can handles only two possible outcome values.

The multinomial logistic regression, however, improves the weakness found in the use of the binary logistic regression. In the multinomial logistic regression analysis, there are rooms for at least three possible outcome values, but these values are not ordered. Vilaça et al. (2019)
use the multinomial logistic regression model to identify statistically significant variables to predict the vulnerable road users risk injuries based on spatial and temporal assessment. Similar studies that used the multinomial logistic regression in RTF include the works of Abdulhafedh (2017), Useche et al. (2018) and Damsere-Derry et al. (2019).

Unlike the multinomial logistic regression, the ordinal logistic regression has its outcome values ordered. The ordinal logistic regression has seen several usages in diverse fields of research (Michalaki et al., 2015; Mehrara, Rezapour & Ksaibati, 2020; Pour-Rouholamin, Jalayer & Zhou, 2017; Rezapour & Ksaibati, 2018). With regards to studies about road traffic fatalities, Kadilar (2016) and Qiana et al. (2020) used the ordinal logistic regression to analyze the effect of drivers’ behavior, roadway and vehicle characteristics on crash severity in traffic accidents in Turkey and China respectively.

Another commonly used probabilistic model in RTF data analysis is the linear regression model. Previous studies from Ghana (Agyemang et al., 2013; Siaw et al., 2013) mainly applied multiple regression analysis to draw conclusion that the road traffic accidents were surging and at a faster rate in Ghana. However, Shankar, Milton and Mannering (1997) critiqued the use of linear regression models as inappropriate for making probabilistic statements about the occurrences of vehicular accidents on the road because it lacks the ability to establish a precise relationship between the dependent variable and independent variables for smaller sample sizes. In such circumstances, Abdullah and Zamri (2012) proposed the use of a fuzzy linear regression models to analyze factors responsible for RTF.

In other related studies (Prasetijo & Musa, 2016; Naji, Xue, Zheng & Lyu, 2020; Sami, Amin, & Butt, 2022), Poisson regression models were fitted for RTF data with the basic assumption that the data produced same mean and variance. However, Shaik and Hossain (2020) faulted the use of the Poisson regression model as in many instances RTF data tends to have a larger variance or over dispersed. For instance, it indicates that Poisson regression model is not suitable for the analysis of both under dispersion and over dispersion data set.

With the knowledge gained in the reviewed literature, we proceed to present statistical methods suitable for predicting RTF in Ghana.

**METHODS**

**Model formulation**

In generalized linear model (GLM), it is assumed that a linear relationship exists between a variable \( Y \), called the dependent variable (outcome variable) and \( k \) independent variables,
Thus, given \( X_1 = x_1, X_2 = x_2 \) and \( X_k = x_k \), the basic model for GLM is of the form

\[
Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon, \hspace{1cm} \text{(1)}
\]

where \( \beta_0, \beta_1, \ldots, \beta_p \) are the partial GLM coefficients, and where it is assumed that \( E(\varepsilon) = 0 \) and \( V(\varepsilon) = \sigma^2 \). The link function, \( g(\mu) \), specifies the link between random and systematic components. It says how the expected value, \( \mu = E(Y) \), of the response relates to the prediction equation of explanatory variables through a prediction equation having linear form

\[
g(\mu) = \eta \hspace{1cm} \text{(2)}
\]

The simplest link function is \( g(\mu) = \mu \). This models the mean directly and is called the identity link. Suppose we want a GLM that models \( n \) Bernoulli trials, that is, the response variable \( Y \) is binary that takes value of 1 (for success) or 0 (for failure). The probability mass function of \( Y \) is given by

\[
P(Y = y) = \pi^y (1-\pi)^{1-y}, \hspace{1cm} y = 0, 1, \hspace{1cm} \text{..................(3)}
\]

where \( \pi \) is the probability of success. Since \( E(Y) = \pi \), it follows that \( \eta \) is a function of \( \pi \). In subsequent sections, we present three different link functions that can be used to estimate \( \pi \).

1. **Logistic regression model**

The link function for logistic regression model is the logit or log-odds function, which is defined, according to McCullagh and Nelder (1989), as:

\[
\eta = \text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right), \hspace{1cm} \text{..................(4)}
\]

Thus, the logistic regression model can be written as

\[
\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \hspace{1cm} \text{..................(5)}
\]

Solving for the probability \( \pi \) in the logit model gives

\[
\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)},
\]

which gives

\[
\pi = \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)}, \hspace{1cm} \text{..................(6)}
\]
where \( \mathbf{x} = (1 \ x_1 \ \cdots \ x_p)' \) and \( \mathbf{\beta} = (\beta_0 \ \beta_1 \ \cdots \ \beta_p)' \). Let \( y_1, y_2, \ldots, y_n \) be the \( n \) observed independent Bernoulli trials with parameters \( \pi_1, \pi_2, \ldots, \pi_n \), respectively. From Equation (3), the likelihood function is given by

\[
\begin{align*}
    l(y|\mathbf{\beta}) &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\
    &= \prod_{i=1}^{n} \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i) \\
    &= \prod_{i=1}^{n} \left( \frac{e^{\mathbf{x}_i'\mathbf{\beta}}}{1 + e^{\mathbf{x}_i'\mathbf{\beta}}} \right)^{y_i} \left( 1 - \frac{e^{\mathbf{x}_i'\mathbf{\beta}}}{1 + e^{\mathbf{x}_i'\mathbf{\beta}}} \right) \\
    &= \prod_{i=1}^{n} \left( e^{\mathbf{x}_i'\mathbf{\beta}} \right)^{y_i} \left( \frac{1}{1 + e^{\mathbf{x}_i'\mathbf{\beta}}} \right) \\
    &= \prod_{i=1}^{n} \left( e^{\mathbf{y}_i\mathbf{x}_i'\mathbf{\beta}} \right) \left( \frac{1}{1 + e^{\mathbf{x}_i'\mathbf{\beta}}} \right). 
\end{align*}
\]

The maximum likelihood estimates of the components of the vector \( \mathbf{\beta} \) are the values of \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) which maximize the likelihood function. They are also the values of \( \beta_0, \beta_1, \ldots, \beta_k \) which maximize

\[
L(\mathbf{\beta}) = \ln l(\mathbf{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i'\mathbf{\beta} - \sum_{i=1}^{n} \ln(1 + e^{\mathbf{x}_i'\mathbf{\beta}}) \]

The first derivative of \( \mathbf{x}_i'\mathbf{\beta} \) with respect to \( \beta_j \) is \( \mathbf{x}_{ij} \), thus:

\[
\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} y_i \mathbf{x}_{ij} - \sum_{i=1}^{n} \left( \frac{e^{\mathbf{x}_i'\mathbf{\beta}}}{1 + e^{\mathbf{x}_i'\mathbf{\beta}}} \right) \mathbf{x}_{ij} = \sum_{i=1}^{n} y_i \mathbf{x}_{ij} - \sum_{i=1}^{n} \pi_i \mathbf{x}_{ij} \\
= \sum_{i=1}^{n} \left( y_i - \pi_i \right) \mathbf{x}_{ij}. 
\]

Setting each partial derivative in Equation (8) to zero, and replacing \( \beta_0, \beta_1, \ldots, \beta_k \) by \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \), we obtain the maximum likelihood estimates of \( \beta_0, \beta_1, \ldots, \beta_k \). The methods of solution are iterative in nature and have been programmed into logistic regression software.
2. **Probit regression model**

The probit regression method uses the cumulative distribution function of the normal distribution to explain the function of the equation. In the probit model, the inverse standard normal distribution of the probability is modeled as a linear combination of the predictors (Fox, 2008; Garson, 2013). Thus, the link function for the model is

\[
\text{probit}(\pi_i) = z_i, \quad \text{..........................................................} \quad (10)
\]

where \( \Phi(z_i) = \pi_i \), and \( \Phi \) is the distribution function of the standard normal distribution. Thus, the probit model has the expression:

\[
\text{probit}(\pi) = x'\beta. \quad \text{..........................................................} \quad (11)
\]

Therefore,

\[
\pi = \Phi(x'\beta). \quad \text{..........................................................} \quad (12)
\]

Therefore, from Equation (3),

\[
P(Y = y) = \left[ \Phi(x'\beta) \right]^y \left[ 1 - \Phi(x'\beta) \right]^{1-y}, \quad y = 0, 1. \quad \text{..........................................................} \quad (13)
\]

Similar to Equation (6), the likelihood function is given by

\[
l(\beta) = \prod_{i=1}^{n} \left[ \Phi(x_i'\beta) \right]^{y_i} \left[ 1 - \Phi(x_i'\beta) \right]^{1-y_i} \quad \text{..........................................................} \quad (14)
\]

The maximum likelihood estimates of \( \beta_0, \beta_1, \ldots, \beta_k \) which maximize the likelihood function. They are also the values of \( \beta_0, \beta_1, \ldots, \beta_k \) which maximize the log-likelihood function. Greene (2011) showed that the estimator \( \hat{\beta} \) could be calculated through maximizing the following log-likelihood function \( L(\beta) \):

\[
\hat{\beta} = \arg \max_{\beta} \left\{ \sum_{i=1}^{n} y_i \ln \Phi(x_i'\beta) + \sum_{i=1}^{n} \left[ (1-y_i) \ln (1-\Phi(x_i'\beta)) \right] \right\}. \quad \text{..........................................................} \quad (15)
\]

3. **Linear probability model**

The link function for linear probability model is the identity, which is defined as \( \eta = \pi \). Thus, the identity regression model can be written as:

\[
\pi = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = x'\beta \quad \text{..........................................................} \quad (16)
\]
The likelihood function is given by
\[
l(y|\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}
\]
\[
= \prod_{i=1}^{n} (x_i^\prime \beta)^{y_i} (1 - x_i^\prime \beta)^{1-y_i} \tag{17}
\]

Log-likelihood function is
\[
L(\beta) = \ln l(\beta)
\]
\[
= \sum_{i=1}^{n} y_i \ln(x_i^\prime \beta) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - x_i^\prime \beta) \ldots \tag{7}
\]

The first derivative of \(x_i^\prime \beta\) with respect to \(\beta_j\) is \(x_{ij}\), thus:
\[
\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i x_{ij}}{x_i^\prime \beta} - \sum_{i=1}^{n} \frac{(1 - y_i) x_{ij}}{1 - x_i^\prime \beta} = \sum_{i=1}^{n} \frac{y_i x_{ij} - x_{ij}^\prime \beta x_{ij}}{x_i^\prime \beta (1 - x_i^\prime \beta)} = \sum_{i=1}^{n} \frac{y_i^\prime - x_{ij}^\prime \beta}{x_i^\prime \beta (1 - x_i^\prime \beta)} x_{ij} \ldots \tag{18}
\]

Setting each partial derivative in Equation (8) to zero, and replacing \(\beta_0, \beta_1, \ldots, \beta_k\) by \(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\), we obtain the maximum likelihood estimates of \(\beta_0, \beta_1, \ldots, \beta_k\).

**The significance of model coefficient**

The hypothesis for testing the significance of any individual regression coefficient such as \(\beta_j\), are
\[
H_0: \beta_j = 0, \quad H_1: \beta_j \neq 0. \tag{19}
\]

If \(H_0: \beta_j = 0\) is rejected, we conclude that, the regressor, \(X_j\), contributes significantly to the model. The Wald chi-squared test statistic for testing \(H_0\) against \(H_1\) is
\[
Z^2 = \left[ \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \right]^2 \tag{20}
\]

where \(\text{se}(\hat{\beta}_j)\) is the estimated standard error of \(\hat{\beta}_j\). When \(H_0\) is true, \(Z^2\) has the chi-square distribution with 1 degree of freedom. We reject \(H_0: \beta_j = 0\) at significance level \(\alpha\) if \(z^2 > \chi_{\alpha,1}^2\).
The comparison of observed to predicted values using the likelihood function is based on the following expression:

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

\[ = 2 \left[ \ln(\text{likelihood of the saturated model}) - \ln(\text{likelihood of the fitted model}) \right] \]........(21)

In particular, to assess the significance of an independent variable, we compare the value of \( D \) with and without the independent variable in the equation. The change in \( D \) due to the inclusion of the independent variable in the model is:

\[ G = D(\text{model without the variable}) - D(\text{model with the variable}) \]

\[ = -2 \ln \left[ \frac{(\text{likelihood of the model without variable})}{(\text{likelihood of the fitted model})} \right]. \]

If the hypothesis that \( \beta_j = 0, \ i = 1, 2, ..., k \) is true, then \( G \) has the chi-square distribution with \( k \) degrees of freedom (Hosmer et al., 1989). It can be shown that a \( 100(1-\alpha)\% \) equal-tails confidence interval for \( \beta_j \) is

\[ \hat{\beta}_j - z_{1-\frac{\alpha}{2}}se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + z_{1-\frac{\alpha}{2}}se(\hat{\beta}_j) \]……………………………………..(22)

**Study Setting**

The National Road Safety Authority (NRSA) in Ghana, which is responsible for ensuring road safety and compliance of road regulation, has recently revealed that a total of 2,924 persons died in 2021 through road crashes. In addition, 15,972 road crashes were recorded within the same period, resulting in 13,048 injuries. We conducted this study in retrospect to past road accidents that occurred from January 2015 to December 2019 in Ghana. Secondary data in the form of recorded road traffic fatalities were obtained from the registry of the NRSA of Ghana for this study. Sixty-six thousand one hundred and fifty-nine (66,159) recorded casualties were included in the current study.

**Data Processing and Analysis**

The secondary data procured from the NRSA were tested for completeness. This was done by first keying all values into Microsoft Excel for cleaning, editing and interpolating for missing values. Thereafter, numerical methods were applied to estimate the proportion of road traffic fatalities across various age groups and gender. The estimation of the models parameters were done using the R statistical application. The respective R programs can be found in appendices A1 and A2.
Table 1 shows the data frame for the logit, probit, and the identity regression models, which is made up of 66,159 casualties who were involve in road traffic accidents in Ghana from 2015 to 2019.

**Table 1: Frequency distribution of Casualties by Age and Gender**

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Gender</th>
<th>Persons Injured</th>
<th>Persons Killed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 5</td>
<td>Male</td>
<td>972</td>
<td>376</td>
<td>1348</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>329</td>
<td>103</td>
<td>432</td>
</tr>
<tr>
<td>6 – 15</td>
<td>Male</td>
<td>2170</td>
<td>563</td>
<td>2733</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>748</td>
<td>153</td>
<td>901</td>
</tr>
<tr>
<td>16– 25</td>
<td>Male</td>
<td>8522</td>
<td>1395</td>
<td>9917</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2890</td>
<td>386</td>
<td>3276</td>
</tr>
<tr>
<td>26– 35</td>
<td>Male</td>
<td>13782</td>
<td>2129</td>
<td>15911</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4737</td>
<td>587</td>
<td>5324</td>
</tr>
<tr>
<td>36– 45</td>
<td>Male</td>
<td>9492</td>
<td>1472</td>
<td>10964</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>3264</td>
<td>406</td>
<td>3670</td>
</tr>
<tr>
<td>46– 55</td>
<td>Male</td>
<td>4245</td>
<td>804</td>
<td>5049</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1461</td>
<td>220</td>
<td>1681</td>
</tr>
<tr>
<td>56– 65</td>
<td>Male</td>
<td>1993</td>
<td>551</td>
<td>2544</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>686</td>
<td>152</td>
<td>838</td>
</tr>
<tr>
<td>Over 65</td>
<td>Male</td>
<td>793</td>
<td>396</td>
<td>1189</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>273</td>
<td>109</td>
<td>382</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>56357</td>
<td>9802</td>
<td>66159</td>
</tr>
</tbody>
</table>

The dependent variable for our models was the number of human deaths per road traffic accident in Ghana, measured on a nominal scale. It is named Casualty ($y$) and coded as Fatality $\equiv$ 1 and Injury $\equiv$ 0. A human death includes passengers, pedestrians, road users and any other person whose death was due to road traffic crush. Two main independent variables were used to assess robustness of our model, namely age groups ($x_1$) and Gender ($x_2$). The age groups were coded according to the increasing order of the fatality index of the age groups. This was necessary because of the use of fatalities per 100 casualties index (F.I.) as a requirement for characterization and comparison of the extent and risk of road traffic fatality (NRSA, 2013; Hesse & Ofosu, 2015). Consequently, age group ($x_1$) was coded as: 0 – 5 years $\equiv$ 7, 6 – 15 years $\equiv$ 5, 16 – 25 years $\equiv$ 3, 26 – 35 years $\equiv$ 1, 36 – 45 years $\equiv$ 2, 46 – 55 years $\equiv$ 4, 56 – 65 years $\equiv$ 6 and over 65 years $\equiv$ 8. Finally, gender ($x_2$) was measured on a nominal scale and coded as: Male $\equiv$ 1 and Female $\equiv$ 2.
Table 2: Rate of fatalities per 100 casualties (fatality indices)

<table>
<thead>
<tr>
<th>Age group</th>
<th>0 – 5</th>
<th>6 – 15</th>
<th>16 – 25</th>
<th>26 – 35</th>
<th>36 – 45</th>
<th>46 – 55</th>
<th>56 – 65</th>
<th>Over 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatality index</td>
<td>26.91</td>
<td>19.70</td>
<td>13.50</td>
<td>12.79</td>
<td>12.83</td>
<td>15.22</td>
<td>20.79</td>
<td>32.15</td>
</tr>
<tr>
<td>Coding</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

RESULTS AND DISCUSSION

Descriptive Statistics

It is observed from Table 1 that the overall male fatalities outnumber their female counterparts by an approximate ratio of 4:1 (78% vs. 21%), for the period 2015 – 2019. Considering the fact that the national population split is slightly in favour of females allude to the fact that male fatalities are highly over-represented in road traffic fatalities. Similar conclusion was reached by Regev et al. (2018) and Melchor et al. (2015), who noted that men are more likely to suffer death in a road accident because of their higher patronage of motor transports for economic activities compared to women.

From Table 2, it can be observed that over the 5-year period, the ‘over 65’ is the age group with the highest national fatality rate. That is, about 32% of all road traffic casualties who were over 65 years lost their lives while 27% of casualties who were 5 years old or less died as a result of road traffic accidents. This finding is consistent with the works of McCoy, Johnston and Duthie (1989), Vaa (2003) and Etehad et al. (2015). Vaa (2003) findings reiterates that people aged 60 years and above may have their biological and psychological system deteriorating faster, and therefore more likely to suffer death in a road accident. Similar explanation could be true for those 5 years and below since they may have a very weak or immature biological and psychological make-up, and therefore more prone to death in case of a road accident. Table 2 shows the fatality indices of each of the eight age groups, computed based on road traffic accident data in Ghana from 2015 to 2019.

Estimated Models

Three models were fitted to the data described in Table 1.0, namely; logistic regression model, probit regression model and linear probability model. Under the logistic regression model, the independent variables were introduced into the model step wisely to measure their individual contribution to the model. In models A, B and C respectively, we mounted the following models:

\[
\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1}, \quad \text{(Model A)}
\]

\[
\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \quad \text{(Model B)}
\]
logit(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}x_{2i} \quad \text{(Model C)}

where \pi_i = P(\text{the } i^{th} \text{ road traffic casualty dies}), x_{1i} \text{ is the age group of the } i^{th} \text{ casualty}, x_{2i} \text{ is the gender of the } i^{th} \text{ casualty and interaction of the variables, } x_{1i}x_{2i}. \text{ The objective is to predict the probability of road traffic fatality as a function of Age group } (x_1) \text{ and Gender } (x_2) \text{ as stated in model B and then compare it with Model A which has only age group as predictor. The analysis using Model B assumes that there is no interaction between Age group and Gender. Therefore, to test for interaction, we compare the full model (Model C) with Model B. The coefficient estimates for the three models and the corresponding standard errors together with their p-values are given in Table 3.}

\textbf{Table 3: Fitted values for the estimated models}

<table>
<thead>
<tr>
<th></th>
<th>Logit Model A</th>
<th>Logit Model B</th>
<th>Logit Model C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>Stand. Error</td>
<td>p-value</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-2.1772</td>
<td>0.0201</td>
<td>0.0000</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.1476</td>
<td>0.0056</td>
<td>0.0000</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.2190</td>
<td>0.0265</td>
<td>0.0000</td>
</tr>
<tr>
<td>( \beta_1 \beta_2 )</td>
<td>-0.0003</td>
<td>0.0135</td>
<td>0.983</td>
</tr>
<tr>
<td>Deviance</td>
<td>54847 on 66157 d.f.</td>
<td>54777 on 66156 d.f.</td>
<td>54777 on 66155 d.f.</td>
</tr>
<tr>
<td>AIC</td>
<td>54851</td>
<td>54783</td>
<td>54785</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Logit Model B</th>
<th>Probit Model D</th>
<th>Identity Model E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>Stand. Error</td>
<td>p-value</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-1.9066</td>
<td>0.0383</td>
<td>0.0000</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.1476</td>
<td>0.0056</td>
<td>0.0000</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.2190</td>
<td>0.0265</td>
<td>0.0000</td>
</tr>
<tr>
<td>Deviance</td>
<td>54777 on 66156 d.f.</td>
<td>54786 on 66156</td>
<td>54830 on 66156</td>
</tr>
<tr>
<td>AIC</td>
<td>54783</td>
<td>54792</td>
<td>54836</td>
</tr>
</tbody>
</table>

Source: Authors Construct (2022), Estimates from R

The results reveal that \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) are significant for all three logistic models in Table 3, with p-values approximating to 0.0000. The observed value of the deviance \( D \) (54847) < \( \chi^2_{0.05, 66157}(66756.45) \) for Model A, which suggests a good fit of the logit regression model. Similar conclusions can be drawn for Models B and C.

The likelihood ratio test for the comparison analysis between Model A and Model B gave a test statistic of 70.2064 with a very small p-value (less than 0.00001). We, therefore,
reject the hypothesis that ‘there is no significant difference between the log-likelihood for Model and that of Model B’, and therefore conclude that the Gender has significant effect in predicting the probability of road traffic fatality. Thus, we would choose to use the **Model B**. This result supports the recent findings of Islam and Mannering (2021) and Useche et al. (2021). Islam and Mannering’s (2021) study confirmed significant differences within gender behavior in predicting the likelihood of a road traffic fatality. Similar conclusion was reached by Useche et al. (2021) whose study support the influence of gender in predicting risky road accidents resulting in deaths.

From Model C, the interaction effect with $p$-value of 0.983 is not significant in predicting the probability of road traffic fatality. Furthermore, a likelihood-ratio test for comparing the model with interaction to model B, without interaction, gave a test statistic of 0.00047 with a $p$-value of 0.9827. Since the $p$-value > 0.05, we conclude that Model C, with interaction, is **not** significantly more accurate than Model B, without interaction. Thus, we would choose to use **Model B**, instead.

In the estimation of the probit regression model, the objective was to predict the probability of road traffic fatality as a function of Age group ($x_1$) and Gender ($x_2$). We then compare it with the logit Model B to determine if there is any significant difference, via likelihood-ratio test. The probit model was formulated as:

$$\text{probit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}. \quad \text{..........................} (\text{Model D})$$

Appendix A3 presents an R function used in estimating the model parameters as given in Table 3. Results from Table 3 shows that the model coefficients are all significant in predicting the probability of road traffic fatality across various age groups in Ghana. Moreover, a residual deviance of 54786 suggests a good fit of the probit model. A comparative analysis of the probit Model D with the logit Model B, via the likelihood ratio test, gave a test statistic value of 9.7934 with a $p$-value of 0.001751, showing that the **Model B** is preferred. Finally, using the linear probability model, we mounted the identity regression model

$$\pi_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}. \quad \text{..........................} (\text{Model E})$$

We fitted the identity regression model using the R code as specified in Appendix A4. The coefficient estimates for Model E and the corresponding standard errors together with the estimates of Model B and Model D are given in Table 3. Like the logit and probit regression models, the linear probability model is significant in predicting road traffic fatality in Ghana.
Comparing the Probit Model E with that of the Logit Model B, via the likelihood ratio test, the value of the test statistic is computed as 53.69624 with a p-value which is less than 0.001, indicating that the logistic regression model is preferred over that of the linear probability model. Table 4 shows the sample proportions of road traffic fatalities and the fitted values for the logistic, probit and linear probability regression models across age groups and gender.

Table 4: Fitted values for the logistic, probit and linear probability regression models

<table>
<thead>
<tr>
<th>Age groups</th>
<th>Gender</th>
<th>Casualties</th>
<th>Proportion Present</th>
<th>Logit fit</th>
<th>Probit fit</th>
<th>Identity Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Persons Injured</td>
<td>Persons Killed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 – 5</td>
<td>Male</td>
<td>972</td>
<td>376</td>
<td>0.2789</td>
<td>0.2511</td>
<td>0.2478</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>329</td>
<td>103</td>
<td>0.2384</td>
<td>0.2122</td>
<td>0.2118</td>
</tr>
<tr>
<td>6 – 15</td>
<td>Male</td>
<td>2170</td>
<td>563</td>
<td>0.2060</td>
<td>0.1998</td>
<td>0.1995</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>748</td>
<td>153</td>
<td>0.1698</td>
<td>0.1671</td>
<td>0.1680</td>
</tr>
<tr>
<td>16 – 25</td>
<td>Male</td>
<td>8522</td>
<td>1395</td>
<td>0.1407</td>
<td>0.1567</td>
<td>0.1574</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2890</td>
<td>386</td>
<td>0.1178</td>
<td>0.1299</td>
<td>0.1306</td>
</tr>
<tr>
<td>26 – 35</td>
<td>Male</td>
<td>13782</td>
<td>2129</td>
<td>0.1338</td>
<td>0.1215</td>
<td>0.1216</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4737</td>
<td>587</td>
<td>0.1103</td>
<td>0.1000</td>
<td>0.0993</td>
</tr>
<tr>
<td>36 – 45</td>
<td>Male</td>
<td>9492</td>
<td>1472</td>
<td>0.1343</td>
<td>0.1382</td>
<td>0.1387</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>3264</td>
<td>406</td>
<td>0.1106</td>
<td>0.1141</td>
<td>0.1142</td>
</tr>
<tr>
<td>46 – 55</td>
<td>Male</td>
<td>4245</td>
<td>804</td>
<td>0.1592</td>
<td>0.1772</td>
<td>0.1777</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1461</td>
<td>220</td>
<td>0.1309</td>
<td>0.1475</td>
<td>0.1485</td>
</tr>
<tr>
<td>56 – 65</td>
<td>Male</td>
<td>1993</td>
<td>551</td>
<td>0.2166</td>
<td>0.2244</td>
<td>0.2229</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>686</td>
<td>152</td>
<td>0.1814</td>
<td>0.1886</td>
<td>0.1891</td>
</tr>
<tr>
<td>Over 65</td>
<td>Male</td>
<td>793</td>
<td>396</td>
<td>0.3331</td>
<td>0.2799</td>
<td>0.2740</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>273</td>
<td>109</td>
<td>0.2853</td>
<td>0.2380</td>
<td>0.2360</td>
</tr>
</tbody>
</table>

The fitted values for the logistic regression model, shown in Table 4, are similar to those for the linear probability and probit regression models. When the values are rounded up to two decimal places, the probit and logistic regression models provide the same estimates.

Conclusion and Recommendation

In this study, we have formulated and estimated three generalized linear models that have shown strong statistical significance in predicting road traffic fatalities given some demographic factors in Ghana. First, the estimated logit regression model proved to be robust when fitted with the predictors, gender and age groupings. However, the interaction of gender and age groupings did not show any significant effect in predicting the probability of road traffic fatality. Secondly, when compared to a probit regression model via a likelihood ratio test, the logit model gave better estimates. Finally, the logit model was preferred over a
fitted linear probability model with the same predictor variables when a likelihood ratio test was conducted to compare the two models. These findings illustrate that a logistic regression model produced consistent probabilities of traffic fatalities which are very close to the actual probability values across age groups and gender. The finding implies that gender and age groups have significant effect in predicting the probability of road traffic fatality. What this means is that the probability of male fatality was high compared to female in a RTA. Also, the aged (over 65 years) were found to be more likely to suffer death in a road accident compared to other categories of age.

Thus, it is recommended to road traffic regulators to consider the possibility of death based on victim’s gender and age group when formulating road safety interventions and regulations to reduce the growing deaths on the road. Road safety education campaigns through the media should be targeted to the male users of transport and the aged, particularly on the use of seat belts in vehicles to reduce the possibility of death in any RTA.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Data Availability Statement

The dataset used in this study is referenced in the paper. However, upon a reasonable request from the corresponding author, the datasets used and/or analysed during the current study could be made available.

Funding

Not applicable

Acknowledgements

Not applicable

References


autoregressive priors. Accident Analysis and Prevention 127, 87-95


Results from meta-analysis (Vol. 690). Oslo, Norway: Institute of Transport Economics.


**APPENDICES**

**Appendix A1**

```r
rtf.logit.Y_X1<-glm(Y~X1, family=binomial(link = "logit"), data=rtf)
rtf.logit.Y_X1_X2<-glm(Y~X1+X2, family=binomial(link = "logit"))
rtf.logit.Y_X1_X2_X1X2<-glm(Y~X1+X2+X1*X2, family=binomial(link = "logit"), data=rtf)
```

**Appendix A2**

```r
A=logLik(rtf.logit.Y_X1)
B=logLik(rtf.logit.Y_X1_X2)
teststat <- -2 * (as.numeric(A)-as.numeric(B))
p.val <- pchisq(teststat, df = 1, lower.tail = FALSE)
p.val
[1] 5.34124e-17
```

**Appendix A3**

```r
rtf.probit.Y_X1_X2<-glm(Y~X1+X2, family=binomial(link = "probit"), data=rtf)
```

**Appendix A4**

```r
rtf.identity.Y_X1_X2<-glm(Y~X1+X2, family=binomial(link = "identity"), data=rtf)
```
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendices.jpg